

生命科学实验指南系列




# Genetic Variation
# A Laboratory Manua

# 遗传变异分析实验指南


［美］M.P. 韦纳　S.B. 加布里埃尔　J.C. 斯蒂芬斯　主编
张根发　谭　信　杨　泽　梁前进　译




科学出版社

(Q-2478.31)

## 生命科学实验指南系列·典藏版




图解微生物实验指南
免疫学技术及其应用
生物衰老：研究方法与实验方案
精编细胞生物学实验指南
植物蛋白质组学实验指南
蛋白质纯化指南（原书第二版）
环境基因组学实验指南
实验动物血液生理生化参考手册
生理学实验指南
精编免疫学实验指南
酵母遗传学方法实验指南
人干细胞培养
抗体制备及使用实验指南
病毒的电子显微学研究
植物生物学与生态学实验
神经生物学实验原理与技术
DNA微阵列实验指南
基因转移：DNA和RNA的转运与表达
生物实验室管理手册（原书第二版）
精编人类遗传学实验指南
单分子技术实验指南
现代蛋白质工程实验指南
活细胞成像（原书第二版）
遗传变异分析实验指南
表皮细胞实验指南
分子克隆实验指南（原书第三版）（上下册）
精编分子生物学实验指南（原书第五版）
现代神经科学研究技术
生命科学实验设计指南
现代生物化学与分子生物学仪器与设备
分子细胞遗传学——技术和应用
精编蛋白质科学实验指南
实验细胞资源的描述标准与管理规范
实验动物设施运行管理指南
元基因组学：方法和步骤（影印版）
现代工业微生物学实验技术
真核生物转录调控——概念策略与技术（原书第二版）
动物细胞培养——基本技术和特殊应用指南（原书第六版）




科学出版中心 生物分社
联系电话：010-64012501
E-mail：lifescience@mail.sciencep.com
网址：http://www.lifescience.com.cn


销售分类建议：分子生物技术



赛拉艾芙
生命科学订阅号


本书彩图请扫码





定价（全套）：4500.00元

“十一五”国家重点图书出版规划项目

生命科学实验指南系列·典藏版

# 遗传变异分析实验指南

## Genetic Variation
## A Laboratory Manual

〔美〕M. P. 韦纳　S. B. 加布里埃尔　J. C. 斯蒂芬斯　主编

张根发　谭　信　杨　泽　梁前进　译

科学出版社

北　京

图字：01-2008-3022 号

## 内 容 简 介

“生命科学实验指南系列”图书均出自名家，包括众多从 Cold Spring Harbor Laboratory Press 和 John Wiley & Sons 等国际知名出版社引进的实验室必备工具书，是生命科学领域最先进、实用、权威的实验手册类优秀图书。该系列图书简单明了，囊括了全世界最著名的生物类实验室操作方法，无论是初学者还是需要深入研究的科研工作者都能从中获益。该系列图书在读者群中有较高的知名度和美誉度，特别是以《分子克隆实验指南》和《精编分子生物学实验指南》为代表，堪称经典，分别被喻为生命科学领域的“蓝宝书”和“红宝书”。现挑选其中的精品集结成典藏版。



图书在版编目（CIP）数据

生命科学实验指南系列：典藏版/雷东锋等编著.—北京：科学出版社，2016
ISBN 978-7-03-047486-5

Ⅰ.①生… Ⅱ.①雷… Ⅲ. ①生命科学–实验–指南 Ⅳ.①Q1-0

中国版本图书馆 CIP 数据核字(2016)第 043878 号

责任编辑：王 静 李 悦
责任印制：张 伟 / 封面设计：刘新新

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京厚诚则铭印刷科技有限公司 印刷
科学出版社发行 各地新华书店经销
*
2016 年 7 月第 一 版 开本：787×1092 1/16
2016 年 7 月第一次印刷 印张：1310 1/2
字数：31 074 000

定价：4500.00 元

(如有印装质量问题，我社负责调换)

# 译 者 序

《遗传变异分析实验指南》一书在 2008 年 6 月刊出的 *Nature Genetics* 有专题推荐介绍。我们团队能够翻译这样一本好书，非常荣幸，也感到责任和压力。为了能更好地将本书展示给读者，出于认真和慎重，本书在我手里已有近两年的时间了。今天，可喜的是这本译著在科学出版社生物分社夏梁编辑的辛勤努力下就要付印了。我大大地松了一口气，真诚地为完成一本能为遗传学研究做出贡献的图书的出版而欣慰。

美国冷泉港实验室被称为世界生命科学的圣地与分子生物学的摇篮，名列世界上影响力最大的十大研究学院榜首。不仅如此，它还是国际生命科学的会议中心与培训基地。冷泉港多兰 DNA 学习中心是全球最有影响的生命科学教育基地，冷泉港实验室出版社因出版名家名著在全球享有盛誉。2008 年初，本书刚刚出版不久，夏梁编辑找到我评价这本书引进、出版的科学意义并征求我对翻译本书的意见。我确实被这本书的新思想、新技术和研究理念打动，因此我放下手头的很多工作，包括有些书籍的撰写和编辑，以高涨的热情和我的同事们一起开始了本书的翻译。由于是新书，很多词汇没有可以借鉴的先例，很多技术和方法也有些生僻，因此为了一个名词的译文我和同事们常常讨论很久，本着为科学不懈努力的精神和不倦追求的态度，一步一步地走来，翻译、校对、再校对地反复工作，我们为减少译著中的错误和使译文更易于理解而尽了最大的努力。功夫不负有心人，在今天本书即将出版之际，我衷心感谢我的同事、朋友和为此书翻译、校对而付出努力工作的学生们，以及为本书出版付出努力的夏梁编辑。

作为翻译本书的首倡者，我召集参与本书翻译的同事们进行讨论分工，规划统一翻译模式，并将全部译稿归纳整理，组织校对和交付出版社编辑之前的工作，同时具体负责了本书的序、前言、第 7～9 章、第 14 章、第 25～27 章及附录等的翻译工作。我的学生李晓峰、刘肖飞、罗雅君参加了部分整理工作。谭信负责第 2～6 章、第 33～35 章的翻译，他的研究生梁雪、孙欢、雷润宏、张公庆参与了校对。杨泽负责本书的第 1 章、第 10～13 章、第 19～24 章的翻译，朱小泉、孙亮、史晓红、万钢、冯洁、张建佚、相蕾参加了部分译校工作。梁前进负责第 15～18 章、第 28～32 章的翻译，王家文、李翠妮、陈哲、罗洋参与了部分翻译工作。参与本书翻译和校对的主要负责人都是本领域的专家，也是遗传学教学一线的骨干，他们扎实的知识和认真负责的科学精神，特别是一丝不苟和踏实认真的科学态度令我深深折服，也鼓励我更多地投入本书的工作。我真心地感谢这些负责人和为此付出努力的同事和学生们。谢谢你们！

毋庸讳言，限于认知水平和知识视野，本书的翻译肯定存有不足乃至错误之处，这应归咎于我的组织和负责不够，敬请读者和相关领域专家指正，以便我有机会时能更好地为遗传学研究著作的引进出版做出自己微薄的贡献。

衷心祝愿我和同事们的成果——《遗传变异分析实验指南》一书能为我国的遗传学研究做出贡献。此乃我辈追求的最高奖赏。

张根发
于北京师范大学
2010年4月16日

# 原 书 序

基因变异的研究从未像现在这样激动人心，这要归功于在连接特殊变异与表型和常见疾病研究方面取得的最新成就。在20世纪的最后10年，人类遗传学团体共同创建了基因连锁图谱，并应用它成功地鉴定了1000多例孟德尔遗传不规则的突变。在21世纪的最初几年迅速绘制的人类基因组序列图谱，使一直处于困境的糖尿病、哮喘、炎症和感染性疾病，以及癌症等常见病的遗传应用初见端倪。人类全基因组序列使大规模地描述人类基因变异的特征成为可能，并使得构建HapMap（国际人类基因组单体型图）计划、对数以百计样品的单一性状基因变异的广泛分析、对数以万计样品的数十万种性状的平行测试和其他新技术的产生达到高潮。人们在其他的有机体，即从微生物到植物和动物付出同等努力，使得连接生物变异与可遗传变异的研究成果一样令人激动。从超越物种的大视角来看，物种内丰富的遗传多样性也同样体现在其表现形式的多样化上，如单核苷酸多态性（SNP）、核苷酸插入或缺失、拷贝数变异、染色质重排和其他结构变异。

在准备这本《遗传变异分析实验指南》时，编辑们组建了一个卓越的团队，为初学者乃至专家打造了坚实的基础。该书的一个醒目特点就是它内容范围的广泛性，其5部分中只有一个部分涉及实验分析的传统领域；而如何运用数据（第2部分）是将来实验研究中常见的实用方法，因此，这部分包含了实验室职员和实习生使用的关键信息。第1部分和第3部分则提供了遗传研究主要方面的全球性知识、集成性研究设计、相关数据库的应用、生物信息学工具、统计分析和解释。同时，也特别涉及了对3种植物（拟南芥、玉米和水稻）和5种哺乳动物（小鼠、大鼠、猫、狗和黑猩猩）感兴趣的研究团队。

人类遗传变异数据库在全球范围的快速增长极大推动了群体遗传学领域的发展和新数据在人类演化史研究中的应用。第3部分则涵盖了一些崭新的方法，这些方法是在研究“自然选择在哪里和怎样在人类进化中塑造现代人的基因组多样性，并且在许多情况下如何导致影响疾病易感性生物途径的一般变异”时被开发出来的。

我提醒学生和其他读者花些时间去阅读该书的介绍、背景、关键概念和该书每一部分的信息，而不仅仅是为了一个具体实验方案或参考某一特殊分析方法才去翻阅这本书。要了解遗传变异，人们需要整合生物学、工艺技术、信息学、数学、统计学和进化生物学。虽然今天的技术必定会像近年来一样继续迅速演变，但是由具有国际水平的专家提出的基本原理将具有持久的价值，帮助你更好地发现新的事实和数据并给予合理的解释。

祝贺编者和作者的卓著工作！

THOMAS J. HUDSON，M. D.
Ontario 癌症研究所

# 前　　言

大概是在2年前，当我们首次着手准备关于基因型研究的实验室指南时，我们的第一想法是，这本指南能够使用多久？就我们2005年底所知的大学和商业实体正在研发的1000美元基因组序列分析技术是否会令基因型分析技术落伍？采纳我们的科学家同事和冷泉港实验室出版社（Cold Spring Harbor Laboratory Press）编辑的建议，我们决定扩展指南的内容，把基因分型合并到完整基因组变异研究的更广大的领域。基于冷泉港实验室一直以来的杰出成果，我们努力想使本书为从事农业和动物领域研究的科研人员提供令他们感兴趣的信息。

在计划这本指南的结构时，我们考虑过给各个不同层次的专业学者提供有用的信息。因此，我们把指南划分成5个部分。第1部分描述研究设计和读者在开始研发之前需要了解的工具。完整的基因组变异数据库的建立令人着迷：你看到的越多，你想要的也越多。但是不管怎样，数据量都不能克服研究计划的不足。实际上，统计学分析表明：一项不漂亮的研究设计中，尽管有可观的基因分型数据也得不到足够的关注，可能产生仅有表面价值的结果，如多重比较设计。第1部分的作者清楚地阐述了这个问题及其相关内容。

虽然很容易想象出所有的基因组研究都从选择基因分型的方法开始，但是我们希望第1部分展示给读者的是一项正确的遗传研究，应从所涉及的问题和恰当的群体选择开始。本指南的第2部分是关于基因分型的基因组变异数据库的建立，共分为3个小部分，主要是提供了人们所期望从冷泉港实验室出版社的实验室指南系列获得的传统“配方和方案”。

第2部分的第1小部分，讲述了从各种不同的样品中分离和制备核酸的方法，包括核糖核酸（RNA）和脱氧核糖核酸（DNA）。第7、8章重点讲述植物遗传研究材料的制备，也包括RNA的分离，因为许多情况不太复杂的转录子组的RNA是比较容易处理的。第9章包括从哺乳动物中制备DNA的12项实验方案，及2个全基因组扩增的2种方法：第一是聚合酶链反应过程，第二是使用“滚圆”复制作用。许多这些方法的试剂盒都可以从商品供应商那里获得，本书提供了试剂盒供应商的地址、配方基本成分和使用中出现问题的解决办法。

第2部分的第2小部分，包括较低和中等通量的SNP基因分型方法。这是基因分型的一种方法（每天可鉴定数百个到一两千个基因型），该方法可保证预期开展一或两项中型设定项目的平均大小的实验室建设。近10年我们看到这种通量的基因分型技术的暴发性增长。但是，这些方法大都可以被视作有2个组分：样品的检测和数据读取。典型的检测包括单一碱基链扩增、TaqMan、寡聚核苷酸连接和等位基因转移性扩增等。读取数据可以使用胶凝、毛细管电泳法、荧光极化和流式细胞仪等完成。

正像DNA测序和oligo-DNA引物的合成，基因分型技术是可替换的。将来，大

多数基于实验室的基因分型工作，可通过外购单独的核心设施来完成，但是对基因分型的产生机理的理解在核定其分型的可靠性工作中仍是重要的。

随着更多基因组拷贝数变异信息的揭示，拷贝数变异成为基因组变异评价越来越高的一个重要方面。发现和测定染色体拷贝数目变化的能力成为我们认识正常状态和疾病状态之间差异的关键因素。幸运的是我们有几个章节涵盖了这个重要的主题。

第 3 部分涉及对获得数据的分析及能帮助我们进行 SNP 选择策略和遗传关联的分析工具。在第 4 部分中，我们设法帮助读者理解生物有机体的当前研究状态，主要包括一些已完成基因组测序和不同水平上分析的几种植物和动物物种。第 5 部分主要描述一些关于人类变异的远景。

本书缺少什么内容呢？我们有意识地放弃了关于超高通量（UHT）基因分型的大量信息。这方面研究的启动是昂贵的，况且技术一直在迅速地发展变化并超出了本书的范围。在我们早期的讨论中，我们曾考虑设一个部分论述能增加对遗传性疾病认识的遗传研究的一个或更多经典实例，如囊性纤维化或糖尿病。取而代之的是我们选择了感觉能引起更广泛兴趣的模式生物。我们也广泛地讨论了关于增加更多不同的模式生物，如果蝇、猪和牛等物种，但是最后，我们认为本书提供的模式生物资料足够使读者理解各种可用的信息。

我们感谢这些使本书计划成为现实的冷泉港实验室出版社的职员。我们赞赏 CSHL 出版社编辑部职员 Inez Sialiano 付出的努力：作为本项目的协调人，是他帮助我们应付与避免自然灾害；感谢自由职业编辑 Irene Pech、Laurie Goodman 和 Martin Winer，特别是 Kaaren Janssen 在项目的创意、排版和管理方面的贡献。我们也感谢出版经理 Denise Weiss、出版编辑 Pat Barker、计算机编辑 Lauren Heller，在本书出版中的专业技术贡献。还要感谢出版社的执行主任 John Inglis，图书发展、市场和营销经理 Jan Argentine 一直对本项目的全力支持和参与。

M. P. Weiner 感谢那些直接或间接帮助和启发他创意的人们，特别是 Camille Solbrig、Evan、Alan Weiner、Albert Wexler、Michael Fleming、Ron Yasbin、Stan Zahler、Ross McIntyre、Ed Buss、Ray Wu、Harold Scheraga、Joe Sorge Jr、Jay Short、Mike Luther、Dan Burns、Lee Babiss、Allen Roses、Phil Whitcome 和 Jonathan Rothberg。

J. C. Stephens 在此感谢那些在他的事业生涯中给予鼓励和指导的所有的同事和朋友们。由于那是一张跨越 30 年时间的冗长的名单，因此他只想选出一些努力提供“大图片”并整合到群体遗传学的良师益友，这些教授有：Wyatt W. Anderson、John Avise、Michael Clegg、Bruce S. Weir、Masatoshi Nei、Wen-Hsiung Li、Ranajit Chakraboty、Frand Ruddle、Ken Kidd 和 Stephen J. O’Brien。

S. B. Gabriel 感谢家庭、朋友和 Broad 研究所的同事们，感谢他们在本书撰写期间给予的不断鼓励和全力支持。

最后，将最崇高的谢意献给我们的作者们，深深感谢他们及时响应我们的建议。H. Leung 博士应该是为本书奉献出不懈努力的最佳代表，尽管在本书创作期间台风肆虐使他所在的研究所停电达一个星期，但他还是准时地完成了有关水稻一章的杰作。同

这些人们一起工作是非常愉快的。

从孟德尔的豌豆到达尔文的鸟雀到 Benzer 的噬菌体再到人类基因组的解密，贯穿着持续的遗传学研究。我们对遗传学遗产和那些用于得出遗传机理的技术的双重理解可以使我们获得认识跨越所有科学的通用“词典”。自从 DNA 双螺旋结构 50 年前被解析以来，它不仅已促使了生物学基础框架的建立，而且已成为所有研究领域的科学家能围绕之交谈的“营火”。我们希望这本指南能让这种交谈继续下去。

MICHAEL P. WEINER
STACEY B. GABRIEL
J. CLAIBORNE STEPHENS

# 目　录

# 1 人类遗传学研究中的伦理学：全球经验

Kevin Arnold and Joelle van der Walt

*Motif BioSciences, New York, New York 10017*



## 简介

人类遗传学研究的飞速发展产生了两个强大而且互相冲突的社会反应。一方面，公众非常期盼人类遗传学的研究能够在医学诊断和治疗方面带来重大突破；另一方面，人们又极为担心人类遗传学的飞速发展可能会增加个人隐私的暴露程度，如个人的遗传基因信息是否会更多地作为法庭的公开数据（Simoncell 2006；Smith 2006）。人们也十分关心如何调整人类遗传学的研究领域使之更好地满足公众利益的需求（ALRC 2003；Manaouil et al. 2005；Swartling et al. 2007）。

人们已经达成大体上一致的意见：为了保护在人类遗传学研究中所涉及的个体的权利，在遗传学的伦理学方面应该遵守 3 个基本原则——尊重个人隐私、尊重个人利益、公平（WHO/CIOMS 2002）。因此，在开展人类遗传学研究之前，必须注意一些重要的伦理学问题，在 2003 年的 ALRC 报告中这些伦理学问题被简单地总结为：

- 哪些人可以被允许去收集、使用或透露某个被收集人或者某些被收集人的遗传信息？
- 谁的个人信息被收集？其个人信息向谁透露？
- 目的何在？
- 得到了谁的同意？

- 以什么样的方式？
- 在什么条件下？

本章接下来的内容总结了一些在人类遗传学领域积极开展研究工作的国际组织和一些国家的实践经验。这些组织和国家曾经限定了遗传学研究在伦理方面的关键范围，并且通过大量的工作，对每一个伦理学的问题进行了详细阐述。本章将要向大家简要介绍目前对于规范人类遗传研究的考虑以及伦理学方面的官方资源和权威机构。这个领域也是一直在向前发展的，如果有时间的话，强烈建议您阅读一下澳大利亚报告（ALRC 2003)，该报告共有1100页内容。基于此报告，您可以非常直接地了解到在遗传学研究中获得、使用、储存以及引申使用人类遗传信息时遇到的伦理学问题。

## 生物医学伦理学的发展简史

在生物医学伦理学发展的历程中，目前绝大多数遗传研究领域的伦理学问题既不是新的也不是唯一的问题。遗传学研究的伦理学框架作为人类生物学和医学的一个研究方面，到目前为止一直在激烈的争论中发展前进。事实上，从第二次世界大战期间医学实验开始发展至今已有60多年，这些伦理学问题一直被激烈地讨论着，人们认识到形成一部伦理学法规是重要和必要的。

表1-1列举了一些生物医学研究中关于伦理学问题的主要历史性文件。表格虽然并非是一个很全面的文件列表，但是可以看出医学研究中的伦理学问题并不是一个新的领域。尽管其中大多数的文件并没有应用于遗传学研究，但是提出的观念已经形成应用于遗传学研究的伦理学思维基础。

**表1-1　生物医学伦理学的简要历史**

| 文件 | 日期 | 注　释 |
|---|---|---|
| 纽伦堡法典 | 1947年 | 第一部被全球认可的医学研究必须遵循的伦理学法规，该法规的制定标志了第二次世界大战的结束<br>http://ohsr.od.nih.gov/nuremberg.html |
| 赫尔辛基宣言 | 1964年<br>(1897年、1983年、1989年修订) | 世界医学联合会(WMA)发展了自己的伦理学指南，对纽伦堡法典的内容进行了扩充<br>http://www.wma.net/e/policy/b3.htm |
| 贝尔蒙报告 | 1979年 | 由美国政府出版，对以前的伦理学法规进行了修订和扩充，使之对人体研究对象的保护达到了一个顶峰<br>http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm |
| 包括人体研究对象的生物医学研究的国际伦理指南 | 1982年<br>(1993年、2002年修订) | 世界卫生组织(World Health Organization,WHO)与国际医学科学组织委员会合作出版，并且进行定期的更新<br>http://www.cioms.ch/frame_guidelines_nov_2002.htm |
| 遗传学研究中的原则声明 | 1996年 | 国际人类基因组组织(Human Genome Organization,HUGO)伦理、法律与社会委员会出版，此声明作为人类基因组计划的一部分<br>http://www.cu.edu.au/eubios/HUGO.html |

续表

| 文件 | 日期 | 注　释 |
| --- | --- | --- |
| 人类基因组和人权的通用声明 | 1997年 | 联合国教育、科学及文化组织(United Nations Educational, Scientific and Cultural Organization, UNESCO)文件,该文件努力体现遗传学研究中的人体研究对象的个人权利<br>http://portal.unesco.org/shs/en/ev.php-URL_ID=1881&URL_DO=DO_TOPIC&URL_SECTION=201.html |
| 人类的本质：澳大利亚人类遗传信息的保护,澳大利亚资源委员会,1996年 | 2003年 | 生物伦理学方面的一个全面性文件,作为规范在澳大利亚获得和使用遗传资源和信息的澳大利亚政策的根据<br>http://www.austlii.edu.au/au/other/alrc/publications/reports/96 |
| 药物遗传学伦理论点 | 2003年 | 英国努菲尔德生物伦理理事会(Nuffield Council on Bioethics)文件<br>http://www.nuffiledbioethics.org/ |
| 人类组织使用法令 | 2004年 | 英国对于器官和组织使用的框架——在英国法律中为了防止"DNA偷窃"行为<br>http://www.opsi.gov.uk/acts2004/2004030.htm |
| 人类生物学材料:收集、使用和储存的推荐 | 2005年 | 爱尔兰生物伦理学委员会的报告文件,推荐用于人类研究中的伦理学行为<br>http://www.bioethics.ie/publications/index.html |

# 伦理学研究的关键

## 知情同意：普遍性

在大家都普遍支持开始遗传学研究之前，应该得到每一个参加研究的对象的知情同意。对于每一个科研项目，研究的组织者都有责任保护每一个受试者和受试家庭的个人隐私，尊重所有人的个人权利。

在如此多的政府代理机构和国际组织中，每一个机构或者组织对于知情同意问题都有其各自的形式。美国关于知情同意所必须遵循的要素的指导方针指定由2家机构代理：卫生与公共服务部（DHHS）——《卫生与公共服务部45号文件》CFR50 part 116/117；食品与药物管理局（FDA）——《食品与药物管理局21号文件》CFR50 part 25。美国的知情同意的指导方针相比世界卫生组织（WHO）和国际医学科学组织委员会（CIOMS）集体合作制定的知情同意的指导方针来说差异较小，所有的知情同意的核心要素是一样的。下面我们对不同的组织和研究机构制定的知情同意原则的一些代表性要求进行描述和比较。

《贝尔蒙报告》由美国政府出版，清楚地限定了知情同意最低所需遵循的三要素：信息、理解和自愿。疾病控制中心主张知情同意应该包含并且传达相应的信息，这些信息可以被每一位可能的研究对象所理解。美国疾病控制中心对于知情同意的指导方针见Box 1-1。

**Box 1-1　美国疾病控制中心：知情同意的要素**

(http://www.cdc.gov/genomics/population/publications/consent.htm)

• 研究名称，并且描述开展该项研究的原因。

• 参加该项研究的参加者需要提供什么？参加者享有什么样的权利？参加者可不可以中途退出该项目？

• 参加者的个人信息是如何被保密的？研究结束时实验样本如何处理？

• 参加该研究的风险/好处是什么？

• 该项研究的参加者需不需要承担任何费用？

• 研究结束后参加者是否被告知结果？

• 参加者在参加研究过程中遇到问题、麻烦或想退出该项研究时应与谁联系？

• 研究的每一个参加者都应该是本人自愿同意参加的。

Box 1-2 介绍了马萨诸塞综合性医院和合作保健网实行的知情同意中必须遵循的可理解要求的主要要素。这些要求来自马萨诸塞综合性医院的知情同意表格中的说明。

**Box 1-2　马萨诸塞综合性医院：知情同意的要素**

(http://healthcare.partners.org/phsirb/confrm.htm)

• 该研究包含研究项目的声明以及研究目的的解释，包括每一个参加者参加该项目的可能持续时间。

• 对于实验程序的描述，如果其中某一个或某些程序是实验性的应该告知参加者。

• 参加者可能遇到的风险或不适的描述。

• 参加者可能获得的利益的描述。

• 如果存在可以选择的治疗程序或方案，应向参加者说明，这对参加者是有好处的。

• 参加者应了解标有受试者身份的试验记录是在什么程度内保密的。

• 对于涉及较大危险的研究，应说明是否给予赔偿和一旦发生损伤可以得到的医学治疗。知情同意书还应说明如果损伤真的发生了，受试者应如何进一步获得有关的信息。

• 明确地告知参加者，如果想进一步了解有关试验的情况或参加者的权利等信息，可与谁联系；如果参加者在试验过程中受到损伤，与谁联系。

• 参加是自愿的。拒绝参加不会有任何过错，也不会缺少受试者所拥有的任何权利；同样，受试者也可以随时退出试验，不会有任何过错，也不会缺少受试者所拥有的任何权利。

**如果合适的话，下列信息要素也应该提供给每一个参加者**

• 特殊的治疗或试验过程可能会对受试者产生危险，而且在目前这又是无法预测的。

- 在什么情况下，不需要受试者的同意，研究者就可以终止试验。
- 如果受试者因为参加了试验而导致额外的费用，研究者应向受试者说明。
- 当退出试验可能对受试者产生危害时，知情同意书应说明为受试者安全起见所必要的退出程序，还要特别地说明这些程序对受试者为什么是重要的。
- 该项研究的可能的研究结果，以及与受试者继续参与该项研究的意愿有关的研究结果。
- 该项研究的受试者的大致数目。

## 知情同意：儿童同意

对于儿童疾病的研究来说，儿童受试者的参与是必不可少的；然而包含儿童受试者的遗传研究的理由目前来说太有限了。为了进一步保护儿童这样一个特殊的群体，相关部门修改了美国联邦关于知情同意的法规，但是其潜在的结果还是不可改变的。简单地说，遗传学研究中的儿童参加者都是在儿童的父母或者监护人的决定下进行的；只能在将来的某个时间，参加试验的儿童长大成人时才具备评估该决定的合法地位。理想的状况是，在一个年龄适合的知情同意下，儿童可以自己决定参加研究。然而，对于所有的知情同意来说，理解研究的意义以及研究结果在将来的应用是高度主观的，因此，想要确保所有儿童的知情同意并不是在父母或者监护人的压力下获得的，是极其困难的(Fisher 2006a，2006b；Panlson 2006)。

我们并不知道儿童试验者的遗传信息将来如何被使用，也不知道儿童试验者的遗传信息的泄漏在将来会不会侵犯该儿童的权利以及该儿童的隐私，因此我们并不清楚将来什么样的规定才是合乎伦理的。因此想要提高参加研究的儿童的权利将来不受侵犯的可能性需要采取谨慎的方案。这种谨慎的方案应该包括，参加试验的儿童一旦具备自己决定自己行为的权利，可以随时选择退出研究，并且可以声明研究中该儿童的遗传信息清楚的状态——是可以自动公开还是不可以公开。

现在已经普遍接受的观点是，任何支持有可能用于儿童的新的治疗和诊断方案的人，对新的治疗和诊断方案在应用之前必须进行评估，评估其应用于儿童的安全性和有效性。在开展有儿童受试者的研究之前，研究人员必须确保以下几点。

- 该项研究不会在成年人中开展得同样有效。
- 研究的目的是获得与儿童健康有关的知识。
- 对每一个儿童受试者的研究都获得了其父亲或者母亲或者监护人的许可。
- 获得了每一个参与研究的儿童本人的同意。
- 必须尊重儿童本人拒绝参与研究或者退出研究的决定。

许可儿童参与研究的父母或者监护人应该给予这样的机会，即在合情合理的范围内能够观察研究的进行情况，这样如果儿童的父母或者监护人认为研究损害到儿童的利益时，可以决定儿童退出研究（WHO/CIOMS 2002)。

## 国际和各国家调控标准及监督者

近年来，美国、英国、澳大利亚、欧洲和中东等国家或地区的政府管理部门和其他一些独立的组织机构对于如何最优地平衡遗传研究中的伦理学和科学问题进行了深入的研究探讨，其结果是这些管理部门和研发机构共同形成了遗传研究的发展框架。

这些组织和机构制定了大量人类遗传学研究所需遵循的指导性文件，这些文件反映了在遗传数据和潜在信息使用上的快速变化。这些指导方针的主要共同点就是，对于遗传研究的开展必须具备理解机构审查委员会（IRB）、科研道德委员会（REC）或者伦理审查机构（ERC）的审查、监督和支持。这三个机构或者委员会有效地作为遗传学研究开展和实施的守门人。

## 已有的单一国家或多国的伦理审查委员会实例

### 英国

在英国，2004 年颁布的英国《人体组织法》（Human Tissue Act）规定，没有本人的同意，储存或者分析任何个人的包含细胞的组织样本都是违法的。当欧盟（EU）颁布了临床试验研究的伦理委员会运行方针时，英国针对此又补充了一系列规定，即《英国关于医学临床试验的人类样本使用的规定，2004》。它规定启动、补充或者实施一项药物临床试验必须得到科研道德委员会（REC）的支持，还要得到英国药物及健康产品管理局（MHRA）的授权，否则被认为非法。

在英国，任何生物医学研究的开展都必须得到国家认可的科研道德伦理委员会的支持，而由英国政府设立且资助的全民健康医疗服务（National Health Service，NHS）体系内的科研道德伦理委员会有 100 多个。这些委员会在工作中遵循标准的操作程序，这个标准的操作程序由科研道德委员会中心办公室撰写（COREC，http://www. corec. org. uk/index. htm），该办公室对科研道德伦理委员会的所有行为负责。科研道德伦理委员会中心办公室（COREC）在英国伦理委员会（UKECA）权威机构的管理下，任何一个科研道德伦理委员会在有权进行临床试验评估之前必须获得英国伦理委员会权威机构的认可。

英国的生物银行与维康信托基金会（Wellcome Trust）、医学研究委员会和卫生部合作设立了自己的伦理管理委员会，形成了其伦理管理的参照标准和准则（2006. 7）（http://www. wellcome. ac. uk/doc _ WTD003284. html）。该参照标准规定：英国生物银行资源的核心科学研究协议、操作程序和资源的使用建议必须得到相关的伦理委员会的支持，必须与相关机构的规定和指导策略相一致；参与者必须被告知该研究会获得独立的伦理学支持；知情同意必须在参与者理解的基础上获得；参与者理解参与的样本在英国生物银行的支持下仅作为研究使用，并且获得了相关的伦理委员会的支持，研究的数据和样本在提供给研究人员之前将被匿名。

英国还主张努菲尔德生物伦理理事会（Nuffield Council on Bioethics）的条例。努菲尔德生物伦理理事会是一个独立的机构，该理事会的唯一功能就是考虑由生物医药领域的新发展而带来的伦理学问题。该理事会曾经专门就在没有正式伦理审查程序的国家

开展科学研究这一问题发表了一篇论文，其论文结论说明：对于这些国家的研究开展而言，尽管研究方案的至少一部分应该由本地的委员会进行审查，审查其是否与当地的健康需求相关，但是利用其他国家的相应机构来完成伦理审查也是可行的。

维康信托基金会资助了一项生物医学伦理计划，以在英国和发展中国家建立伦理学的知识基础并且发展研究的接受能力。作为一个资助机构，维康信托基金会对于申请者编制了财务状况表和指导性说明。这个文件声明了申请者在有人类受试者参与的研究中的地位：对提交的研究方案如何通过道德审查委员会的审查负有责任（http://www.wellcome.ac.uk/node5240.html）。然而维康信托基金会也接受这样的情况，如果该国家没有道德审查委员会，但是研究还要开展，那么可以选择其他的道德审查方式，在这种情况下，维康信托基金会可以将其申请转至伦理学常委咨询委员会（SAGE）来进行建议。

### 澳大利亚

澳大利亚，2001年发动了一项全面的调查，调查结果就是在2003年出版了一份1100页的全面报告——*Essentially Yours：The Protection of Genetic Information in Australia*（ALRC 2003）。该报告认为1998年的个人隐私法已经包含了对个人遗传信息的保护，因此并不需要制定关于伦理审查的新的法律规定。尽管该报告导致包括人类受试者的研究的道德审查行为需由澳大利亚国家医学研究委员会（NHMRC）进行，并且形成了一个新的法定机构对人类遗传学研究中出现的伦理学问题向政府提供建议，但是为了符合国际观点，对于任何特殊的遗传研究的监督还是通过科研道德委员会完成的。

该报告承认已经存在的国际指导方针，并且认为如《2003年世界人类基因组和人权宣言》这样的准则应该被考虑参照（www.unesco.org/ethics）。

同其他国家一样，由澳大利亚国家健康和医学研究委员会发布的“国家声明”规定：包括人类受试者的研究方案必须要经过人类科研道德委员会（HREC）的审查和批准；人类科研道德委员会的主要功能就是保护研究中受试者的权利。“国家声明”和人类科研道德委员会的另一个作用是帮助那些对研究人员的小群体或者对整个人类有利的研究顺利进行，这个作用经常被忽视。

### 美国

在美国，美国食品与药物管理局对遗传研究的开展进行有效的管理，同英国的管理方式相似，要求任何包含人类受试者的研究都必须得到理解机构审查委员会的批准。美国食品与药物管理局为理解机构审查委员会公开发表了指导方针。例如，理解机构审查委员会的功能在21CFR part 56（http://www.fda.gov/oc/ohrt/irbs/appendixc.html）。在美国，所有包含人类受试者的生物医学研究都要求有理解机构审查委员会的正式批准，理解机构审查委员会的审核根据美国食品与药物管理局和卫生与公共服务部的条文规定。

美国国立卫生院（NIH）负责管理和资助生物医学领域中大型的政府科研预算，美国国立卫生院规定：“没有理解机构审查委员会的批准的研究是不会被资助的。”相似地，美国食品与药物管理局要求制药公司收集人体的遗传数据以及使用其产生的遗传数据都要经过理解机构审查委员会的批准。

美国食品与药物管理局承认，在没有关于美国食品与药物管理局如何通过批准的程序使用这些遗传数据的政策的情况下，制药公司可能并不情愿去执行由美国食品与药物管理局控制的包含药物基因组学成分的临床试验。因此，2005 年 3 月，美国食品与药物管理局公布了药物基因组学资料提交、允许自愿提交基因组数据的工业指导准则（VGDS）（http://www.fda.gov/cber/gdlns/pharmdtasab.htm）。美国食品与药物管理局指出如果在药物基因组学检测中存在有侵害性的检测，包括抽血，都必须在研究方案和知情同意书中清楚地注明。这是美国食品与药物管理局在促进遗传学研究、评价个体由遗传因素引起的对药物的反应方面作出的努力。

瑞典

瑞典，研究中伦理学参照标准方面也经历了很多的变化，目前在研究的伦理学方面采用欧盟的规范文件。瑞典 2004 年颁布了包括人类受试者研究的伦理审查的法律文件，该法律文件更新了审查的研究程序，指定了中央和地方的新的审查委员会结构，限定了地方性的审查委员会的组成。

在瑞典，瑞典研究委员会对于研究中伦理学的参照标准制定负有主要的责任。瑞典研究委员会与卡罗琳斯卡医学院和乌普萨拉大学的生物伦理学中心合作制定了 CODEX（http://www.codex.vr.se/codex_eng/codex/index.html）：科学研究中伦理方面的可理解信息的网站。在 2003 年伦理审查法律条文中，瑞典政府就推荐瑞典研究委员会负责任命除中央伦理审查委员会之外的所有伦理审查委员会的成员，中央伦理审查委员会的成员由政府直接任命。

国际人类基因组单体型图联盟

在过去的 5 年中，国际人类基因组单体型图联盟（International HapMap Consortium）是在世界范围内成功解决遗传研究中伦理学问题的复杂的研究项目之一。国际人类基因组单体型图联盟是一个多中心的、多国家的研究团体，其中包括加拿大、中国、日本、尼日利亚、英国和美国的许多科学家和资助机构。建立国际人类基因组单体型图联盟的目的是获得一个公共资源库来帮助研究者找到与人类健康和疾病相关的基因。国际人类基因组单体型图联盟成功地由参与国家的理解机构审查委员会网络管理（International Hap Map Consortium 2004）。

## 关于伦理学可接受的遗传学研究的最新思考

遗传学研究中的伦理学问题是一个不断发展的领域，从目前的考虑来看，一直由理解机构审查委员会和科研道德委员会来负责的伦理学问题正在远离无限制接受的、模糊的、不明确规定的科研项目，而越来越靠近那些在各方面都非常明确清晰的科研项目。目前，所有包含遗传研究的项目计划都要明确说明其关注的疾病、将要研究的基因和（或）通过全基因组分析所能获得的遗传信息，以及研究样本的存储时间。

遗传学研究中伦理学问题的不断发展，并不是由单独的审查委员会引起的，而是各国或者全球的组织对于遗传研究中伦理学的、法律的以及社会的含义等问题考虑的结

果。这些意见主要是基于公众的或者商业资助的大规模生物资源库的潜力以及考虑到试验样本已经获得的遗传信息及其将来可能的使用情况而得出。在美国，1990 年国家人类基因组研究所（NHGRI）设立了伦理、法律和社会含义（ELSI）研究计划（http://www.genome.gov/policyethics），该计划作为人类基因组计划不可缺少的成分，促进遗传和基因组研究在伦理、法律和社会含义方面的进展。

### 知情同意的限制

UNESCO、WHO、COREC，澳大利亚的 NHMRC、EU、HUGO-ELSI 和其他组织出台了许多遗传研究中的伦理学审查指导方针（准则）。然而，目前的资料还规定了生物伦理学中知情同意和试验结果公开的标准（Clayton 2005；Deschenes and Sallee 2005；Hoeyer et al. 2005）。由于这些指导方针和标准是不断更新的，被修改过的文件很可能包括对于知情同意或分层的知情同意的限制、对再次知情同意的需要、对登记的受试个体研究结果公开的新要求、对发展中国家伦理审查能力的促进等（Hyder et al. 2004；Bhat and Hegde 2006；Gbadegsin and Wendler 2006）。

全面的知情同意允许研究者在目前的研究中没有限制地保留试验样品，并不要求研究者在将来的研究中对于试验样品的使用再次获得受试者的知情同意，并且受试者不可以或者只可以有限制地退出研究。尽管对于样本的研究使用是匿名的，对于知情同意的约束力还是仅能在特殊的情况下被接受。例如，国际人类基因组单体型图联盟和英国的生物银行计划。理解机构审查委员会希望知情同意的形式和方案能够被清楚地定义和限定范围——并不是限制研究人员，而是给参与研究的受试者以更清晰的解释。显然，从包含一个较全面的无预期目标的知情同意的项目转变成一个关注特定疾病的项目（有特定的遗传界限和特定的时间范围），可以使理解机构审查委员会对该项目有更清晰的监督。

显然，如果回复到无预期目标、不明确规定的项目设计上，那么主框架将会有很大地移位，从个体权利到更多集权方面。如果匿名参加项目研究，也就是说研究中实验样本是匿名的，这种个人权利的丧失看起来会减轻。然而，匿名参加研究会破坏（或者严重地限制）受试者退出该项研究的能力，并且一旦匿名，审查委员会对于匿名样本的遗传检测的限制就会放宽。然而，将来理解机构审查委员会还是会要求受试者个体保留退出研究的权利。使人担心的是，受试者的退出会限制这些完全匿名的样本的使用。考虑到当前分子遗传学技术发展，受试者的退出是可以实现的。希望退出研究的受试者仅需再次匿名提供个人样本，声明收回与该样本匹配的所有数据即可。

## 结论

总之，这些围绕遗传学研究的伦理学问题并不新鲜，也不是唯一的。在人类生物学研究中伦理学考虑的范围也一直在发展；然而，多数意见仍认为对于伦理学研究来说需要确定关键的观念思维。任何包含人类受试者参加的科学研究计划，在开始实施之前都应该得到科研道德伦理委员会（或者理解机构审查委员会、伦理审查机构等）的审查和批准。尽管不同国家的科研道德伦理委员会的成员资格和成员人数有很大的差别，但国

家以及国际的伦理学审查指导方针一贯认可这些机构的审查和监督。

从参加者的观点来看，所有的准则方针都很直接地要求在任何研究开展之前都应该获得所有受试者的知情同意，知情同意也必须是建立在每一个参加研究的人自愿、知情、理解的基础上。

# 参考文献


Australian Law Reform Commission (ALRC). 2003. *Essentially yours: The protection of human genetic information in Australia*, report 96. Australian Law Reform Commission, Sydney. http://www.austlii.edu.au/au/other/alrc/publications/reports/96/

Belmont Report. 1979. *Ethical principles and guidelines for the protection of human subjects of research*. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. U.S. Department of Health, Education, and Welfare, Washington, D.C. http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm

Bhat S.B. and Hegde T.T. 2006. Ethical international research on human subjects research in the absence of local institutional review boards. *J. Med. Ethics* **32:** 535–536.

Clayton E.W. 2005. Informed consent and biobanks. *J. Law Med. Ethics* **33:** 15–21.

Corrigan O.P. and William-Jones B. 2006. Pharmacogenetics: The bioethical problem of DNA investment banking. *Stud. Hist. Philos. Biol. Biomed. Sci.* **37:** 550–565.

Deschenes M. and Sallee C. 2005. Accountability in population biobanking: Comparative approaches. *J. Law Med. Ethics* **33:** 40–53.

Fisher C.B. 2006a. Privacy and ethics in pediatric environmental health research—Part II: Protecting families and communities. *Environ. Health Perspect.* **114:** 1622–1625.

———. 2006b. Privacy and ethics in pediatric environmental health research—Part I: Genetic and prenatal testing. *Environ. Health Perspect.* **114:** 1617–1621.

Gbadegesin S. and Wendler D. 2006. Protecting communities in health research from exploitation. *Bioethics* **20:** 248–253.

Hansson M.G., Dillner J., Bartram C.R., Carlson J.A., and Helgesson G. 2006. Should donors be allowed to give broad consent to future biobank research? *Lancet Oncol.* **7:** 266–269.

Hoeyer K., Olofsson B.O., Mjorndal T., and Lynoe N. 2005. The ethics of research using biobanks: Reason to question the importance attributed to informed consent. *Arch. Intern. Med.* **165:** 97–100.

Hyder A.A., Wali S.A., Khan A.N., Teoh N.B., Kass N.E., and Dawson L. 2004. Ethical review of health research: A perspective from developing country researchers. *J. Med. Ethics* **30:** 68–72.

International HapMap Consortium 2004. Integrating ethics and science in the International HapMap project. *Nat. Rev. Genet.* **5:** 467–475.

Knoppers B.M. 2005. Biobanking: International norms. *J. Law Med. Ethics* **33:** 7–14.

Manaouil C., Graser M., Chatelain D., and Jarde O. 2005. The examination of genetic characteristics since the adoption of the French law on bioethics. *Med. Law* **24:** 783–789.

Paulson J.A. 2006. An exploration of ethical issues in research in children's health and the environment. *Environ. Health Perspect.* **114:** 1603–1608.

Rothstein M.A. 2005. Expanding the ethical analysis of biobanks. *J. Law Med. Ethics* **33:** 89–101.

Shickle D. 2006. The consent problem within DNA biobanks. *Stud. Hist. Philos. Biol. Biomed. Sci.* **37:** 503–519.

Simoncelli T. 2006. Dangerous excursions: The case against expanding the forensic DNA databases to innocent persons. *J. Law Med. Ethics* **34:** 390–397.

Smith M.E. 2006. Let's make the DNA identification database as inclusive as possible. *J. Law Med. Ethics* **34:** 385–389.

Swartling U., Eriksson S., Ludvigsson J., and Helgesson G. 2007. Concern, pressure and lack of knowledge affect choice of not wanting to know high-risk status. *Eur. J. Hum. Genet.* **15:** 556–562.

WHO/CIOMS 2002. *International ethical guidelines for biomedical research involving human subjects.* Council for International organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO), Geneva. http://www.cioms.ch/frame_guidelines_nov_2002.htm


# 互联网资源

**http://healthcare.partners.org/phsirb/consfrm.htm** Consent forms. Partners HealthCare System, Partners Human Research Committee.

**http://www.cdc.gov/genomics/population/publications/consent.htm** Informed consent template for population-based research involving genetics. National Office of Public Health Genomics, Department of Health and Human Services, Centers for Disease Control, Atlanta, Georgia.

**http://www.codex.vr.se/codex_eng/codex/index.html** CODEX rules and guidelines for research homepage. CODEX is The Swedish Research Council's gateway to various research ethics guidelines. In collaboration with The Centre for Bioethics at Karolinska Institute and Uppsala University.

**http://www.corec.org.uk/index.htm** Central Office for Research Ethics Committees (COREC) homepage. National Patient Safety Agency, London, United Kingdom.

**http://www.fda.gov/oc/ohrt/irbs/appendixc.html** Information Sheets. Guidance for Institutional Review Boards and Clinical Investigators 1988 Update. U.S. Food and Drug Administration, Rockville, Maryland.

**http://www.fda.gov/cber/gdlns/pharmdtasub.htm** Guidance for Industry page. Pharmacogenomic Data Submissions. U.S. Food and Drug Administration, Rockville, Maryland.

**http://www.genome.gov/PolicyEthics/** Policy & Ethics homepage. National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland.

**http://www.wellcome.ac.uk/doc_WTD003284.html** U.K. Biobank Ethics and Governance Frameword homepage. Wellcome Trust, London, United Kingdom.

**http://www.wellcome.ac.uk/node5240.html** Biomedical ethics. Reports on meetings and other activities organized by the Wellcome Trust Biomedical Ethics Programme, London, United Kingdom.

**www.unesco.org/ethics**

# 2 遗传分析的群体选择

J. Claiborne Stephens[1] and Michael Bamshad[2]

[1]*Motif BioSciences Inc., New York, New York 10017;* [2]*Departments of Pediatrics and Genome Sciences, Division of Genetics and Developmental Medicine, University of Washington School of Medicine, Seattle, Washington 98195*



## 简介

遗传关联和基因作图研究主要由 5 个部分组成：研究设计、表型分型、样品的采集和处理、基因型分型、统计分析。本书后面的一些章节讲到了后三部分，而表型分型则大大超过了本书的范围。对于研究设计，主要的考虑包括对象是家庭还是无关的个体、统计学所需要样品的大小、特定群体是否可以在科学上有用等。由于现在已经清楚选择适宜的遗传学研究材料时群体的特性非常重要，因此本章将重点放在后者（Jorde 1995；de la Chapelle and Wright 1998；Wright et al. 1999）。

有几种类型的群体可以对遗传分析提供特殊便利。利用隔离的群体或"建立者"群体进行人类遗传学研究已经有很长的历史，且结果丰富。部分原因是在"建立者"群体中许多疾病的流行程度较普通群体高或低很多，因而致病基因的鉴定相对容易。这种流行程度的不同是因为"建立者"群体都经历过一个人口规模的压缩阶段，称为"瓶颈"，而这一小群体只携带原有群体遗传变异的一部分。因此致病基因在这个小的"建立者"群体中或者快速地流行开来，或者趋于灭亡。

与普通群体相比，非随机婚配的群体类型对遗传学研究也更有价值。其中，在亲属之间的血亲婚姻（如在一级表亲之间）是首先选择的婚姻类型。由于血亲婚姻，罕见的常染色体隐性遗传病得以显露出来。纯合性或"同接合性"（autozygosity）作图是致病基因作图的一种特殊策略，血亲婚姻后代出现基因组 DNA 的长同接合性地带的可能性增加，这一点可被利用。同接合性本质上是 DNA 的纯合性，造成这种纯合性的原因是

一个基因组区域的两个 DNA 拷贝均来自最近的共同祖先。这一方法的另一个常用的术语是“世系的同一性”作图（“identity-by-descent” mapping）。

最近混合不平衡作图或者简单的“混合作图”已被用来定位很多在医学上有重要价值的基因。混合群体是指那些在历史上曾经分开而又重新融合或再次相遇的群体。利用混合群体进行作图的原因是在最初的祖先群体中任何有较大基因频率差异的遗传性状在混合群体中都相对容易。

在本章中，我们讨论“建立者”群体、纯合性作图、混合群体，并且考虑它们在人类致病基因作图方面各自的优势。

## “建立者”群体

“建立者”群体和建立者效应的观念已经深深地植入群体遗传学中。在人类中，一个“建立者”群体是指从一个较大的群体中分离出来的由数目很少的个体发展起来的群体。随着时间的推移，这个新的群体保持着与其他相关群体的隔离状态，通过随机过程产生了与其他群体不同的基因频率。由于历史上的地理隔离，许多岛屿群体都是“建立者”群体。遗传隔离的另一个来源是对相互通婚的限制，这种限制可能有宗教的、语言的、社会的或文化上的原因。

在阐明许多孟德尔式疾病的遗传基础时，人类“建立者”群体已经被证明具有极大的价值。例如，当致病等位基因的携带者由于机会原因在“建立者”群体中比来源群体更常见时，就会暴露出许多罕见的常染色体隐性遗传病。许多家庭会带有本来罕见的疾病，而且对这一疾病来说这些家庭具有典型的共同遗传基础，这使得遗传基础的判断十分容易。

最近已普遍认识到，“建立者”群体受限的遗传变异性有利于一般的复杂疾病的遗传研究（Service et al. 2006；Freimer and Sabatti 2007）。除了彼此相似的遗传背景外，“建立者”群体中的个体常生活在相同的环境下。对于一个常见的复杂疾病，一个生活在广大的、背景多变的大陆上的个体的易感性可能与几十种遗传或环境因素有关，不难想象，一个“建立者”群体中这样的遗传或环境的干扰因素将减少。因此，如果一个常见复杂疾病在大陆群体和“建立者”群体之间有发病率差异的话，可预计在“建立者”群体中具有更多导致疾病潜在变异的遗传基础。如果确实如此，在“建立者”群体中存在一个或少数对于疾病易感性十分重要的基因的可能性会大大增加。鉴定一个或多个显效基因显然比鉴定许多微效基因容易得多。

“建立者”群体的遗传结构也有利于致病基因的鉴定和作图。遗传结构中特别重要的部分是连锁不平衡（LD）的程度和类型。LD 是一个群体的个体之间遗传标记共同出现的倾向性。正是这种共同性，使得我们可以取一组标记，如微阵列中的一组单核苷酸多态性（SNP），来推论整个基因组哪个区域与特定的表型有关或无关。目前已经显示出在许多“建立者”群体中 LD 的水平较一般群体中特异性高，而整个 LD 类型的差异比一般群体中的小（Service et al. 2006）。

## "建立者"群体：实际应用

对隔离群体大量出色的研究已经阐明了一些重要的孟德尔式疾病的遗传基础。一些群体因其社团组织、参与和所开展的遗传工作而闻名。Ashkenazi 犹太群体（Risch et al. 1995；最近的综述请看 Charrow 2004；Weinstein 2007）具有许多组织良好的网站和遗传监控系统，监控至少 9 种疾病［Tay-Sachs 病、Canavan、高雪病（Gaucher）、Niemann-Pick 病、家族性自主神经异常（familial dysautonomia）、布卢姆综合征（Bloom syndrome）、1A 型糖原贮积病、范可尼贫血（Fanconi anemia）、黏脂贮积病Ⅳ型（Mucolipidosis IV）］，这些疾病在这个群体中发病率异常的高。一些常见病，如乳癌和囊性纤维化病，在 Ashkenazi 群体中也有独特的或突出的高频率变异体（Quint et al. 2005；Chen et al. 2006；Weinstein 2007）。

另一个开展与其群体规模不相称的遗传学研究的群体是芬兰群体（de la Chapelle and Wright 1998；Peltonen et al. 1999；Peltonen 2000；Kere 2001）。在芬兰，一些特殊的常染色体显性遗传病或隐性遗传病有较高的发病率，推测有"建立者"效应（请参见 Kere 2001 文献的表 1）。通过广泛的登记和活跃的疾病遗传基础研究工作，芬兰的遗传学研究已经深入到对致病基因的鉴定，以及众多的遗传病基础的系统研究中。

在美国，对一些由于宗教信仰的原因而与大众隔离的社团的研究已对基因作图研究产生重大影响。Victor McKusick 对 Old Order Amish（阿们宗派，是 17 世纪成立的一个门诺教派，因创此教派的雅各·阿们而得名，其生活方式与主流社会脱离，译者注）展开的工作是最早期的工作之一，它清楚地表明某些疾病具有遗传基础（McKusick et al. 1964；McKusick 1973）。这个广大的家族对遗传方式的估计（判断是显性的还是隐性的）提供了方便。在能应用遗传标记将致病基因定位于特定染色体之前，许多出色的工作就已经完成了。一旦遗传标记可以使用，许多在 Old Order Amish 中流行的疾病的染色体基因定位工作就可以开始（Velinov et al. 1993；Polymeropoulos et al. 1996）。

最近在哈特派信徒（Hutterites）中的哮喘研究（Chan et al. 2006；Kurz et al. 2006）表明在基因组的许多区域存在风险变异。宾夕法尼亚东南部的门诺派教徒社团（Mennonite，16 世纪起源于荷兰的基督教新派，译者注）是研究某些在其他地方罕见的遗传疾病的对象，如 Hirschsprung 症和槭糖尿病（Puffenberger 2003）。后者在门诺派教徒中的发病率比随机婚配的欧洲群体高 1000 倍。

对委内瑞拉马拉开波湾湖（Lake Maracaibo）的隔离群体的亨廷顿病（Huntington disease，HD）研究是一个十分重要的隔离群体的遗传学研究事例。尽管 HD 是早期基因作图成功的事例之一，其基因在 20 世纪 80 年代中期就被定位在 4 号染色体短臂上（Gusella et al. 1983），但又经过了差不多 10 年时间科学家才确定了具体的基因（MacDonald et al. 1992）。这个委内瑞拉群体差不多有 100 位 HD 患者，他们在最终的作图和基因的确定上起着关键作用。

最近 Bonnen 等对一个遗传结构非常简明的隔离群体的研究（2006）是非常有趣的例子。他们评估了对科斯雷（Kosrae）密克罗尼西亚岛进行全基因组相关研究的可能

性。他们使用全基因组范围的 SNP 资料（30 个核心家系的 113 240 个 SNP），发现与 HapMap（International HapMap Consortium，国际人类基因组单体型图联盟）中同样的 SNP 基因型相比，该岛 LD 增加且等位基因多样性减少。他们进一步发现 LD 的长度比 HapMap 中的长。作者认为科斯雷群体所作的贡献在于该群体与其他群体相比具有遗传学研究的有效性。

## 纯合性作图

在 1987 年的一篇里程碑式的论文中，Eric Lander 和 David Botstein 提出 DNA 过度纯合的受累儿童可以用于进行常染色体隐性致病基因的基因组定位（Lander and Botstein 1987）。特别是通过近亲婚配出生的少量受累儿童和相对较少的基因组标记（就目前的标准而言）就能达到定位疾病或常染色体隐性遗传的性状基因的目的。其原则出自这样一种思想：从近婚群体而来的受累儿童对同一个决定祖传性状的等位基因来讲是完全纯合的，决定这个性状的基因位于每个受累儿童的纯合性区域之内。在每个儿童的基因组中预期会有许多纯合性区域，但只有所有受累儿童共有的区域才可能是引起性状的假定区域。进一步来讲，实际的区域应该限定在所有受累儿童纯合性区域的重叠区，并排除任何非受累同胞的纯合性区域（对于同一套等位基因）。

现在已经通过纯合性作图在众多的群体中进行了几十个常染色体隐性遗传病的作图（Sheffield et al. 1998）。这样的群体中有许多倾向于一级表亲婚姻或其他近亲婚姻，因此具有较高的罕见疾病发病率（Teebi and El-Shanti 2006）。其他群体可能避免近亲结婚，但也可能由于“建立者”效应或者地理隔离及其他隔离效应而增加血亲婚配的机会（Strauss et al. 2005）。

## 混合群体

50 多年前，成百上千的遗传标记尚不可被利用，Rife（1954）的一篇充满洞察力的文章就提出混合群体在致病基因作图方面可能存在价值。在混合群体中，个体的祖先来自两个或多个祖先群体，这些祖先群体在此前彼此隔离。34 年后，Chakraborty 和 Weiss（1988）在一篇理论文章中描绘了对致病基因作图的策略，使这一思想得到进一步创新和发展。按照这一策略，任何在各个来源群体中流行程度不同的疾病或性状应该显示出对于连锁标记的 LD，在不同来源群体之间这些连锁标记也会有频率差异。这一策略的明显优点是与性状连锁的标记并不需要与性状位点本身靠得很近。因此混合 LD 作图，或者更简单的混合作图（admixture mapping）所需的标记数目比起更常规的研究少了很多。

这一策略隐含着几个假定。首先，所探讨的性状应具有遗传学基础，并且这些性状在来源群体中存在程度的差异应该与有关遗传因子在来源群体中的频率差异有关。其次，混合群体初次形成后应经过足够的时间，以便使任何性状和非连锁的标记之间的 LD 趋于消失。这不能排除正在进行的与任意一个来源群体混合的可能性，但在混合的

进程中，与非连锁的标记相比，连锁产生的信号将明显增加。注意，群体形成后的时间、祖先群体的相对贡献和混合过程动力学是一些应考虑的重要因素，它们决定了在给定群体中使用这一策略进行致病基因作图的可行性。

从 20 世纪 90 年代的中期到末期，众多论文对这一作图方法提出了更多的可行方法（Stephens et al. 1994；McKeigue 1997，1998；Shriver et al. 1997）。此后不久，以整个基因组范围为资源进行混合作图成为可能（Dean et al. 1994；Smith et al. 2001，2004；Collins-Schramm et al. 2002），最终形成一套 4222 个已定位的、记录良好的、高信息量的、已确认的 SNP。这项研究现已能够用于非裔美国人群体（Tian et al. 2006）。毋庸置疑，在更多标记被鉴定和评估后，随着更多的混合群体被研究，这样的标记数还将增加。

当这些标记可以利用后，几项实验研究肯定了理论的预期。其中一项研究（Lautenberger et al. 2000）显示非裔美国人在 Duffy 血型座位（FY）和标记之间有长达 30cM的延伸的 LD。另一项研究（Collins-Schramm et al. 2003）也显示非裔美国人群体中包括成百万的碱基长度区域的 LD。

除了已明确定性的各组标记资源外，理论的和分析的资源也已齐头并进，已发展出几种用于混合作图的特殊算法：MALDSOFT（Montana and Pritchard 2004）、ANCESTRYMAP（Patterson et al. 2004）和 ADMIXMAP（Hoggart et al. 2004）等。最近 Tang 等（2006）介绍了马尔科夫-隐马尔科夫模型（Markov-hidden Markov model，MHMM）算法，用于重建混合群体遗传家系的各环节（SABER，http://www.fhcrc.org/science/labs/tang/）。

在最近的几项研究中，对致病基因使用混合作图的效能已经显现出来。第一个成功的运用是对非裔美国人群微卫星的扫描，结果两个涉及高血压易感性的基因组区域被鉴定出来（Darvasi and Shifman 2005；Zhu et al. 2005），所使用的标记数目（269 个）和受试者数目（试验组加对照组：1310）都相对较小，而且其中一个区域此前已经被提示与高血压有关，这给混合作图方法的有效性提供了支持。

第二个混合作图例子成功地利用了多发性硬化（multiple sclerosis，MS）群体在流行上的差别。多发性硬化被认为是高度遗传的，但还没有令人信服地确定候选基因。由于北欧起源的人群比非洲起源的人群多发性硬化的发病率高，对非裔美国人中的多发性硬化患者使用混合作图可以潜在地鉴定欧洲来源的 MS 易感性基因。在对 1648 例非裔美国人的患者对照研究中，确实发现了位于 1 号染色体上的高度显著性的易感座位（Reich et al. 2005）。

第三个成功事例是对非裔美国男性的前列腺癌的易感性使用混合作图（Freedman et al. 2006）。流行病学上，年龄较轻的非裔美国男性比其他群体来源的同龄男性患前列腺癌的风险显著升高，这使前列腺癌成为应用混合作图策略进行研究的重要目标。连锁分析也提示了同样的基因组区域，8q24，但连锁分析结果只是部分地解释了混合信号的强度。作者正确地预见了位于 8q24 的多个座位涉及易感性。最近，包括他们自己在内的几个实验室已证明了这一点（Gudmundsson et al. 2007；Haiman et al. 2007；Yeager et al. 2007）。

到目前为止，混合作图的着重点在非裔美国人群。然而可以清楚地看到，利用这些人群对多种疾病的成功阐明将激发标记系列的创新，并促进科学家在其他群体中开展研究。

## 结论

如果遗传学研究的焦点是在特殊的性状上，那么对该性状是否可以在一个特殊群体中更方便地进行研究评估就具有很大的意义。由于一个性状存在于某群体中就“从最近的地方开始入手”(looking under the lamppost) 也许不是研究一个复杂性状的遗传基础最有效的途径。在对感兴趣的疾病的遗传特征进行研究时，与研究者自身的群体或国家相比，特殊的外来群体可能更合适。正如我们所认识到的，遗传学研究正在成为更加全球化的事业，除了少数例外，我们都拥有同样的基因，所以在这个星球的任何地方得到的遗传学的发现都可以外推到全人类群体。既然我们从人类遗传学研究的全球化中得到了那么多，我们不久很可能将看到如上面所述的研究在更广泛的人类群体中重复和流行。

## 参考文献


Bonnen P.E., Pe'er I., Plenge R.M., Salit J., Lowe J.K., Shapero M.H., Lifton R.P., Breslow J.L., Daly M.J., Reich D.E., et al. 2006. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat. Genet.* **38:** 214–217.

Chakraborty R. and Weiss K.M. 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci.* **85:** 9119–9123.

Chan A., Newman D.L., Shon A.M., Schneider D.H., Kuldanek S., and Ober C. 2006. Variation in the type I interferon gene cluster on 9p21 influences susceptibility to asthma and atopy. *Genes Immun.* **7:** 169–178.

Charrow J. 2004. Ashkenazi Jewish genetic disorders. *Fam. Cancer* **3:** 201–206.

Chen S., Iversen E.S., Friebel T., Finkelstein D., Weber B.L., Eisen A., Peterson L.E., Schildkraut J.M., Isaacs C., Peshkin B.N., et al. 2006. Characterization of *BRCA1* and *BRCA2* mutations in a large United States sample. *J. Clin. Oncol.* **24:** 863–871.

Collins-Schramm H.E., Chima B., Operario D.J., Criswell L.A., and Seldin M.F. 2003. Markers informative for ancestry demonstrate consistent megabase-length linkage disequilibrium in the African American population. *Hum. Genet.* **113:** 211–219.

Collins-Schramm H.E., Phillips C.M., Operario D.J., Lee J.S., Weber J.L., Hanson R.L., Knowler W.C., Cooper R., Li H., and Seldin M.F. 2002. Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am. J. Hum. Genet.* **70:** 737–750.

Darvasi A. and Shifman S. 2005. The beauty of admixture. *Nat. Genet.* **37:** 118–119.

Dean M., Stephens J.C., Winkler C., Lomb D.A., Ramsburg M., Boaze R., Stewart C., Charbonneau L., Goldman D., Albaugh B.J., et al. 1994. Polymorphic admixture typing in human ethnic populations. *Am. J. Hum. Genet.* **55:** 788–808.

de la Chapelle A. and Wright F.A. 1998. Linkage disequilibrium mapping in isolated populations: The example of Finland revisited. *Proc. Natl. Acad. Sci.* **95:** 12416–12423.

Freedman M.L., Haiman C.A., Patterson N., McDonald G.J., Tandon A., Waliszewska A., Penney K., Steen R.G., Ardlie K., John E.M., et al. 2006. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci.* **103:** 14068–14073.

Freimer N.B. and Sabatti C. 2007. Human genetics: Variants in common diseases. *Nature* **445:** 828–830.

Gudmundsson J., Sulem P., Manolescu A., Amundadottir L.T., Gudbjartsson D., Helgason A., Rafnar T., Bergthorsson J.T., Agnarsson B.A., Baker A., et al. 2007. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39:** 631–637.

Gusella J.F., Wexler N.S., Conneally P.M., Naylor S.L., Anderson M.A., Tanzi R.E., Watkins P.C., Ottina K., Wallace M.R., Sakaguchi A.Y., et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306:** 234–238.

Haiman C.A., Patterson N., Freedman M.L., Myers S.R., Pike M.C., Waliszewska A., Neubauer J., Tandon A., Schirmer C., McDonald G.J., et al. 2007. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39:** 638–644.

Hoggart C.J., Shriver M.D., Kittles R.A., Clayton D.G., and McKeigue P.M. 2004. Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.* **74:** 965–978.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Jorde L.B. 1995. Linkage disequilibrium as a gene-mapping tool. *Am. J. Hum. Genet.* **56:** 11–14.

Kere J. 2001. Human population genetics: Lessons from Finland. *Annu. Rev. Genomics Hum. Genet.* **2:** 103–128.

Kurz T., Hoffjan S., Hayes M.G., Schneider D., Nicolae R., Heinzmann A., Jerkic S.P., Parry R., Cox N.J., Deichmann K.A., and Ober C. 2006. Fine mapping and positional candidate studies on chromosome 5p13 identify multiple asthma susceptibility loci. *J. Allergy Clin. Immunol.* **118:** 396–402.

Lander E.S. and Botstein D. 1987. Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children. *Science* **236:** 1567–1570.

Lautenberger J.A., Stephens J.C., O'Brien S.J., and Smith M.W.

2000. Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. *Am. J. Hum. Genet.* **66:** 969–978.

MacDonald M.E., Novelletto A., Lin C., Tagle D., Barnes G., Bates G., Taylor S., Allitto B., Altherr M., Myers R., et al. 1992. The Huntington's disease candidate region exhibits many different haplotypes. *Nat. Genet.* **1:** 99–103.

McKeigue P.M. 1997. Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *Am. J. Hum. Genet.* **60:** 188–196.

———. 1998. Mapping genes that underlie ethnic differences in disease risk: Methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am. J. Hum. Genet.* **63:** 241–251.

McKusick V.A. 1973. Genetic studies in American inbred populations with particular reference to the Old Order Amish. *Isr. J. Med. Sci.* **9:** 1276–1284.

McKusick V.A., Hostetler J.A., Egeland J.A., and Eldridge R. 1964. The distribution of certain genes in the Old Order Amish. *Cold Spring Harbor Symp. Quant. Biol.* **29:** 99–114.

Montana G. and Pritchard J.K. 2004. Statistical tests for admixture mapping with case-control and cases-only data. *Am. J. Hum. Genet.* **75:** 771–789.

Patterson N., Hattangadi N., Lane B., Lohmueller K.E., Hafler D.A., Oksenberg J.R., Hauser S.L., Smith M.W., O'Brien S.J., Altshuler D., et al. 2004. Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74:** 979–1000.

Peltonen L. 2000. Positional cloning of disease genes: Advantages of genetic isolates. *Hum. Hered.* **50:** 66–75.

Peltonen L., Jalanko A., and Varilo T. 1999. Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.* **8:** 1913–1923.

Polymeropoulos M.H., Ide S.E., Wright M., Goodship J., Weissenbach J., Pyeritz R.E., Da Silva E.O., Ortiz De Luna R.I., and Francomano C.A. 1996. The gene for the Ellis-van Creveld syndrome is located on chromosome 4p16. *Genomics* **35:** 1–5.

Puffenberger E.G. 2003. Genetic heritage of the Old Order Mennonites of southeastern Pennsylvania. *Am. J. Med. Genet. C Semin. Med. Genet.* **121:** 18–31.

Quint A., Lerer I., Sagi M., and Abeliovich D. 2005. Mutation spectrum in Jewish cystic fibrosis patients in Israel: Implication to carrier screening. *Am. J. Med. Genet. A* **136:** 246–248.

Reich D., Patterson N., De Jager P.L., McDonald G.J., Waliszewska A., Tandon A., Lincoln R.R., DeLoa C., Fruhan S.A., Cabre P., et al. 2005. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat. Genet.* **37:** 1113–1118.

Rife D.C. 1954. Populations of hybrid origin as source material for the detection of linkage. *Am. J. Hum. Genet.* **6:** 26–33.

Risch N., de Leon D., Ozelius L., Kramer P., Almasy L., Singer B., Fahn S., Breakefield X., and Bressman S. 1995. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat. Genet.* **9:** 152–159.

Service S., DeYoung J., Karayiorgou M., Roos J.L., Pretorious H., Bedoya G., Ospina J., Ruiz-Linares A., Macedo A., Palha J.A., et al. 2006. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat. Genet.* **38:** 556–560.

Sheffield V.C., Stone E.M., and Carmi R. 1998. Use of isolated inbred human populations for identification of disease genes. *Trends Genet.* **14:** 391–396.

Shriver M.D., Smith M.W., Jin L., Marcini A., Akey J.M., Deka R., and Ferrell R.E. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.* **60:** 957–964.

Smith M.W., Lautenberger J.A., Shin H.D., Chretien J.P., Shrestha S., Gilbert D.A., and O'Brien S.J. 2001. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. *Am. J. Hum. Genet.* **69:** 1080–1094.

Smith M.W., Patterson N., Lautenberger J.A., Truelove A.L., McDonald G.J., Waliszewska A., Kessing B.D., Malasky M.J., Scafe C., Le E., et al. 2004. A high-density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* **74:** 1001–1013.

Stephens J.C., Briscoe D., and O'Brien SJ. 1994. Mapping by admixture linkage disequilibrium in human populations: Limits and guidelines. *Am. J. Hum. Genet.* **55:** 809–824.

Strauss K.A., Puffenberger E.G., Craig D.W., Panganiban C.B., Lee A.M., Hu-Lince D., Stephan D.A., and Morton D.H. 2005. Genome-wide SNP arrays as a diagnostic tool: Clinical description, genetic mapping, and molecular characterization of Salla disease in an Old Order Mennonite population. *Am. J. Hum. Genet.* **138:** 262–267.

Tang H., Coram M., Wang P., Zhu X., and Risch N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* **79:** 1–12.

Teebi A.S. and El-Shanti H.I. 2006. Consanguinity: Implications for practice, research, and policy. *Lancet* **367:** 970–971.

Tian C., Hinds D.A., Shigeta R., Kittles R., Ballinger D.G., and Seldin M.F. 2006. A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. *Am. J. Hum. Genet.* **79:** 640–649.

Velinov M., Sarfarazi M., Young K., Hodes M.E., Conneally P.M., Jackson C.E., and Tsipouras P. 1993. Limb-girdle muscular dystrophy is closely linked to the fibrillin locus on chromosome 15. *Connect. Tissue Res.* **29:** 13–21.

Weinstein LB. 2007. Selected genetic disorders affecting Ashkenazi Jewish families. *Fam. Comm. Health* **30:** 50–62.

Wright A.F., Carothers A.D., and Pirastu M. 1999. Population choice in mapping genes for complex diseases. *Nat. Genet.* **23:** 397–404.

Yeager M., Orr N., Hayes R.B., Jacobs K.B., Kraft P., Wacholder S., Minichiello M.J., Fearnhead P., Yu K., Chatterjee N., et al. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39:** 645–649.

Zhu X., Luke A., Cooper R.S., Quertermous T., Hanis C., Mosley T., Gu C.C., Tang H., Rao D.C., Risch N., and Weder A. 2005. Admixture mapping for hypertension loci with genome-scan markers. *Nat. Genet.* **37:** 177–181.

# 互联网资源

**http://www.fhcrc.org/science/labs/tang/** SABER: SNP-based Ancestry Block Estimation and Reconstruction program. Fred Hutchinson Cancer Research Center, Seattle, Washington (see Tang et al. 2006).

# 3 功效计算

David M. Evans[1] and Shaun Purcell[2]

*[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; [2]Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114*



## 简介

统计测试效能是指它能够产生统计学上的显著结果，说明无效假说不成立的可能性。换句话说，它表示研究能成功地发现真实效应的机会，这依赖于一系列因素，包括效应的大小、样品的大小和研究的设计，以及特定的假阳性率。效能计算在研究的计划阶段开始，通常在确定需要的样品量时着手进行；有时也在对已经结束的研究结果说明时考虑统计效能，特别是在对阴性结果的解释方面。本章我们将回顾统计效能的基础，讨论其能力是怎样估计的（使用遗传的事例/对照研究作为例子），并考虑影响遗传研究能力的最相关因素。最后我们聚焦于在现代全基因组相关研究背景下的效能，其中覆盖、多重测试、阶段设计等问题是非常重要的。

为使效能的基础坚实，考虑如下情况：一个以检测遗传的事例/对照与一个功能的位点（一个位点确实涉及发病风险）的关联为目的的研究。我们从一些假定的群体重复取样，得到每个样品的基因型和表型，计算相关比值比（odds ratio），然后对比值比是否与 1.0 不同（是否拒绝没有相关假设的无效）进行显著性检验。由于随机采样过程，

每个重复样品估计比值比会稍微有所不同。由于取样误差，测试的有些部分是显著的，有些则不是。确切的比值比依赖于因素的量，包括真正作用的大小，因此所有的重复可能都是重要的。从中重复地获得显著性结果的比例表示对测试效能的估计，它告诉我们如果该研究在完全相同的条件下再重复一次，其成功的可能性有多大。实际上我们并不是通过多次试验计算效能，而是依赖于分析理论，或使用计算机模拟实施重复的系列试验。效能计算提出可能的最小研究规模，该研究仍能有效地检测出所有达到一定程度的效应。

对统计效能的考虑（发现真实效应的机会）马上使人关注两类推论性的错误：没有发现真实的效应（Ⅱ型错误）和发现了不存在的效应（Ⅰ型错误）。一个“假的阳性”或Ⅱ型错误对应于未能拒绝无效假设（得到非显著性试验结果），此时相反的假设是真实的（特殊的非无效的相反的假设事实上是真的）。犯Ⅱ型错误的可能性，用希腊字母 $\beta$ 表示，等于 1 减去统计效能：如果相反假设是真，我们或者发现了效能（概率是 $1-\beta$），或者未能发现它（概率是 $\beta$）。

反之，如果无效假设事实上是真的，那么我们或者正确地不拒绝它（该实验不显著），或不正确地拒绝无效（如果试验是显著性的，概率是 $\alpha$），通常试验者通过决定使用什么阈值来决定实验显著性来控制 $\alpha$。用这一方式可以描述和控制一个统计实验的灵敏性和特异性。很自然在两种性质之间有一个平衡：降低阈值增加灵敏性（增加效能，减少Ⅱ型错误）但也降低特异性（增加了Ⅰ型错误率）。历史上许多小规模的研究中，研究者按惯例采纳像 $\alpha=0.05$ 和 $\beta=0.20$（80%效能）的值，表示一个现实和合适的平衡。

在图 3-1 中表示这些值之间的关系（为了方便显示，我们假定作显著性的单侧检验）。图 3-1 显示在无效假设（左边）和相反假设（右边）下的检验统计量 $X$ 的分布。垂直线代表测试的临界值，在此线右边的值被认为是显著的，而左边不显著。在临界值右边的相反假设区域下的值代表检测效能（黑色区和灰色区），而临界值左边的相反假

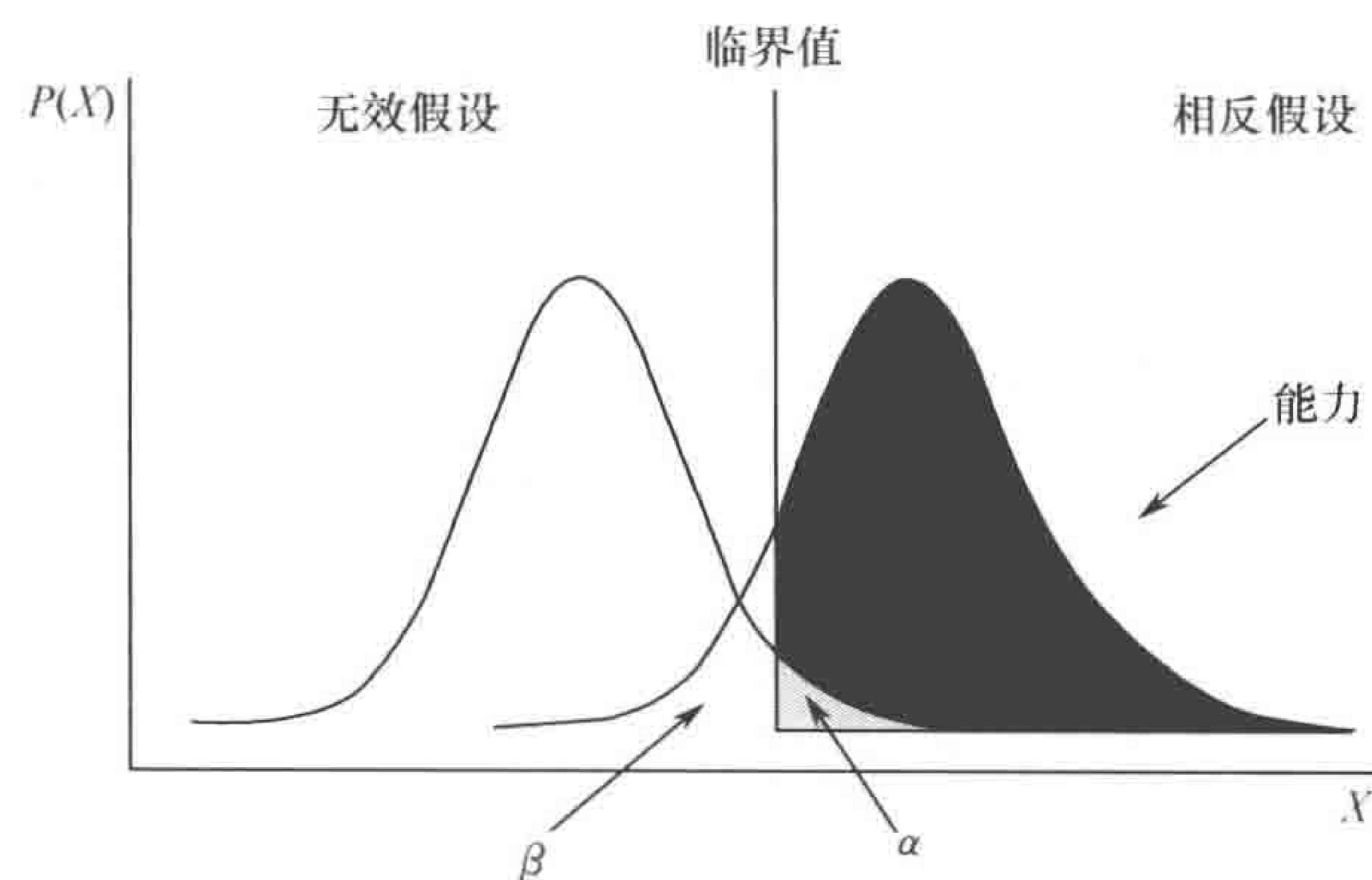



图 3-1　无效（左侧）和相反（右侧）假设的检验统计值的分布。垂直线代表测试的临界值。在相反假设区域下的黑色和灰色的区域代表检测效能。遮蔽的灰色区域表示检验的Ⅰ型错误率，在相反假设区域下的未被遮蔽的区域是犯Ⅱ型错误的可能性

设代表犯Ⅱ型错误的可能性。犯Ⅰ型错误的可能性由在临界值右边的无效假设区域内的灰色区域表示。

除了改变临界值外，还有两种其他方式增加统计测试的效能：增加样品的量和增加效应量。如下所述，期待的测试统计值是样品量和效应量的函数。样品量通常由研究者控制；不过效应量也可以增加虽然并不明显。例如，可通过使用更精确的表型确定减少测量误差，或通过对感兴趣的区域作更精密的基因型分型做到这一点。

如果我们知道在无效和相反的假设下测试统计值的分布，就可能计算出统计学测试分析的期待效能（Sham et al. 2000）。例如，许多遗传关联测试服从卡方分布。卡方分布的形状由“非中心参数”（non-centrality parameter，NCP）决定。当无效假设为真，NCP 为零，形成的卡方分布称为中心卡方分布。在相反的假设下，NCP 将大于零，形成的卡方分布称为非中心卡方分布。这意味着在相反假设下的测试统计值分布改变了形状，根据 NCP 所量化的值的大小移向右侧。期待的测试统计值等于 NCP 加上自由度。由于中心和非中心卡方分布的形状已知，如果我们知道期待的 NCP、测试的自由度和使用的临界值，就有可能推断出在图 3-1 曲线下的区域（以及效能）。

## 遗传模型参数

进行把握度分析前，研究者必须事先决定一些参数的取值。这包括已知的和需要研究者控制的量（如样品的大小、Ⅰ型错误率、疾病的流行等）。研究者也必须对未知的参数作出假设，包括遗传效应的大小和潜在的遗传方式、在标记和位点之间的连锁不平衡程度、在这些位点的基因频率等。举例来讲，相对风险参数 $R_{AA}$ 和 $R_{Aa}$ 量化了感兴趣基因的效应量，表示一个基因型为 $AA$ 的个体（或基因型为 $Aa$ 的个体）患病可能比低风险的 $aa$ 基因型个体（随意假设 $aa$ 是低风险的）高出的倍数。只要有了这些参数和疾病的流行知识（$K$），就可能计算一个 $aa$ 基因型个体患病的绝对风险（$g_{aa}$）。

$$g_{aa}=K/(P_A^2R_{AA}+2p_Ap_aR_{Aa}+p_a^2)$$

一般认为，常见的复杂疾病不可能是显效基因座作用的结果。基因型的相对危险性更可能在 1～1.5。在数量表型的情况下，效应大小的量度与由数量性状基因座产生的全部表型变异成比例。通常该值在很小的百分比范围内（0～5%），只有当有证据表明存在主要基因座时，我们才可以假定有大的效应量。

可用相对风险参数 $R_{AA}$ 和 $R_{Aa}$ 之间的关系推测潜在的遗传模式。例如，$R_{AA}$ 的相对风险=4、$R_{Aa}=1$ 提示 $A$ 等位基因的隐性遗传方式，这是因为 $Aa$ 杂合子没有导致比隐性纯合子更多的风险；反之，$R_{AA}$ 的相对风险=16、$R_{Aa}=4$ 暗示遗传的积乘方式，因为 $AA$ 基因型的风险是杂合子风险的平方。

分型的标记与造成疾病的功能变异可能无关，而仅是与真正的功能基因座处于连锁不平衡的一个标记。性状和标记位点连锁不平衡的程度可以通过 Lewontin's（1964）$D'$系数加以量化。$D'=1$ 值提示标记和疾病基因座之间在历史上未发生过重组，因此是完全的不平衡；反之，$D'=0$ 提示标记和性状基因座之间彼此独立。如果对一个小的基因座区域分型，或对一系列候选基因都感兴趣，可以假定至少一个标记与推定的功能变

异处于高度的连锁不平衡中（当然，实际上假定一个常见变异位于这一区域）。不过由于现在可利用的商业产品在基因组范围捕捉常见变异的效能还不完善，用这一假定在基因组范围进行关联研究还不切实际。例如，最近估计 Affymetrix 111 K、500K 芯片和 Illumina HumanHap300 panel 通过连锁不平衡各自捕获大约 31%、65%和 75%在西欧群体中常见的遗传变异（Barrett and Cardon 2006）。

最后，有必要限定在性状和标记位点的基因频率。除非研究者只对频率推定为已知的特定 SNP 感兴趣，他通常在可能的基因频率范围内进行效能计算。如果使用特异标记常见变异的商业产品（如 Illumina 300K panel of markers），由于这些商品的标签 SNP 不能很好地捕捉罕见变异（Barrett and Cardon 2006），可能必须假定一个“常见”疾病等位基因的存在（较低的等位基因频率>5%）；反之，如果研究的焦点是非同义的 SNP（即 SNP 在蛋白质中产生一个氨基酸的变化），研究者可以假定引起疾病的基因频率更小（由于非同义 SNP 倾向于较低的等位基因频率）。

## 通过分析发现关联的计算效能

在对潜在的遗传模型进行了假设后就有可能计算预期的 NCP 并估算遗传关联试验的能力。几种关联测试的 NCP 的闭式表达（closed-form expression）已经在文献中有所描述，包括传输不平衡测试（TDT；McGinnis et al. 2002）、遗传相关的事例/对照测试（Schork 2002）、数量性状不平衡测试（Sham et al. 2000）。使用这些公式，研究者在各种不同的情景下可以快速和容易地确定效能，不必进行耗费时间的数据模拟。此外，与检查相关能力不同，统计测试的 NCP 本身就是一个有用的量，它随着样品量增加呈线性增加（Witte et al. 2000）。有一系列用具可以帮助研究者在多种不同的研究设计中计算发现遗传关联的期待能力，包括 QUANTO（http://hydra.usc.edu/GxE）、power for association with error（PAWE；http://linkage.rockefeller.edu/pawe）和 Purcell's on-line Genetic Power Calculator（http://pngu.mgh.harvard.edu/~purcell/gpc/）（Gauderman 2002；Gordon et al. 2002；Purcell et al. 2003）。

为阐明关联遗传测试的效能如何能通过分析计算出来，我们来描述为等位关联的 Pearson 卡方检验所做的计算 NCP 的步骤。等位关联的 Pearson 卡方检验比较了事例和对照的标记等位基因的频率。

$$\chi^2 = \sum_{i=0,1} \sum_{j=A,U} \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]}$$

这里合计覆盖了全部等位基因（$i$）和疾病状态（$j$），$n_{ij}$ 是带有疾病状态（$j$）的个体中含有所观察的 $i$ 型等位基因数目的数值，而

$$E[n_{ij}] = \frac{n_{i.}\, n_{.j}}{n_{..}}$$

是在给与边际的共计下（total at the margins）期待的数值。这个等位基因的检测前提是遗传积乘的方法和哈代-温伯格平衡对一般群体限定（对一个个体的两个等位基因依据独立的观察进行处理是恰当的）的认定。这就是说，该测试将基因型分解形成 2×2

可能性表，计数与个体相对的等位基因。如果该认定不成立，关联的基因型测试或 Armitage 倾向测试可能更合适（Sasieni 1997）。

为了阐明用于关联的等位基因测试的 NCP 的来源，考虑一个单一的造成疾病的双等位基因座，上面有等位基因 $A$ 和 $a$，有 $n$ 个事例和 $rn$ 个对照（对照与事例的比为 $r$）。我们指定由 $K$ 引起疾病的流行，高风险等位基因频率为 $P_A$，$Aa$ 和 $AA$ 基因型的相对风险分别为是 $R_{Aa}$ 和 $R_{AA}$。给予上述变量后，就很容易计算一个具有低风险基因型 $aa$ 的个体受染风险。

$$g_{aa} = \frac{K}{R_{AA}p_A^2 + 2R_{Aa}p_A(1-p_A) + (1-p_A)^2}$$

因此，基因型为 $Aa$ 和 $AA$ 的个体患病的绝对风险为

$$g_{Aa} = R_{Aa}g_{aa}$$

$$g_{AA} = R_{AA}g_{aa}$$

从这些数据可以用 Bayes 定理计算事例中各基因型的推测比例。

$$c_{aa} = \frac{g_{aa}(1-p_A)^2}{K}$$

$$c_{Aa} = \frac{2g_{Aa}p_A(1-p_A)}{K}$$

$$c_{AA} = \frac{g_{AA}p_A^2}{K}$$

同样，在对照中

$$u_{aa} = \frac{(1-g_{aa})(1-p_A)^2}{1-K}$$

$$u_{Aa} = \frac{2(1-g_{Aa})p_A(1-p_A)}{1-K}$$

$$u_{AA} = \frac{(1-g_{AA})p_A^2}{1-K}$$

现在可简单计算出在事例（$c_A$）和对照（$u_A$）样品中期待的基因频率：

$$c_A = c_{AA} = \frac{c_{Aa}}{2}$$

和

$$u_A = u_{AA} + \frac{u_{Aa}}{2}$$

这些预期的基因频率列在可能性表（表 3-1）中。

**表 3-1　事例/对照等位基因关联测试的预期基因频率**

| 等位基因 | 事例 | 对照 |
|---|---|---|
| $A$ | $2C_An$ | $2u_Arn$ |
| $a$ | $2(1-C_A)n$ | $2(1-u_A)rn$ |
|  | $2n$ | $2rn$ |

现在可以使用这些值根据 Mitra（1958）首先采用的方法计算 NCP（用希腊字母 $\lambda$

表示）

$$\lambda = \frac{2nr(c_A - u_A)^2(1+r)}{(c_r + ru_A)(1+r-c_A-ru_A)}$$

通过在非中心卡方分布下的整合得到等位基因关联测试的效能。

$$1-\beta = \int_{\chi'^2_{\alpha}(\nu,0)}^{\infty} d\chi'^2(\nu,\lambda)$$

式中，$\chi'^2_{\alpha}(\nu,0)$为自由度为$\nu$的中心$\chi^2$分布的百分点；100％(1－$\alpha$)，$\chi'^2(\nu,\lambda)$为用非中心参数$\lambda$得到的非中心$\chi^2$分布。

该基本步骤假定研究者有幸得到实际性状位点的基因型。然而这种情况在实际中并不常见，除非试图复制一个已知的关联。因此重要的是要扩展这些运算以适合性状位点没有进行基因型分型，而是存在连锁不平衡标记的情况。考虑一个双等位标记位点$M$和$m$，其对应的频率为$p_M$和$p_m$。我们用 Lewontin（1964）的$D'$系数量化在性状和标记位点之间的连锁不平衡。

$$D' = \frac{p_{AM} - p_A p_M}{\min(p_A p_m, p_a p_M)} = \frac{\delta}{\delta_{\max}},\ \delta > 0$$

和

$$D' = \frac{p_{AM} - p_A p_M}{\min(p_A p_M, p_a p_m)} = \frac{\delta}{\delta_{\max}},\ \delta < 0$$

式中，$p_{AM}$为$A$和$M$等位基因单倍型的频率；$\delta$量化了观察的和（独立的）预期的单倍型频率的差别。$\delta>0$时，$\delta_{\max}$是在标记和疾病位点特定基因频率给予$\delta$的可能的最大正值；$\delta<0$时，$\delta_{\max}$是$\delta$可能的最大负值（Devlin and Risch 1995；McGinnis et al. 2002）。在表 3-2 中标记和性状位点的单倍型频率用$\delta$来说明。

**表 3-2　在标记和疾病位点的单倍型**

| | | 性状等位基因 | |
|---|---|---|---|
| | | $A$ | $a$ |
| 标记 | $M$ | $p_A p_M + \delta$ | $p_a p_M - \delta$ |
| 等位基因 | $m$ | $p_A p_m - \delta$ | $p_a p_m + \delta$ |

为了将连锁不平衡整合进 NCP 的运算，我们只要确定与每个标记基因型相关的风险即可。这可以通过将性状位点的风险与标记基因型给予的性状基因型的可能性相乘，然后将所有可能的性状基因型求和来获得。举例来讲，一个在标记位点为$MM$基因型的个体受染的可能性按如下方法运算：

$$P(D=\text{感染的}\,|\,G_M = MM) = \sum_{G_T} P(D=\text{感染的}\,|\,G_T)P(G_T\,|\,MM)$$

$$= \sum_{G_T} P(D=\text{感染的}\,|\,G_T)\frac{P(G_T, G_M = MM)}{P(G_M = MM)}$$

$$= \frac{g_{AA}(p_A p_M + \delta)^2 + 2g_{Aa}(p_A p_M + \delta)(p_a p_M - \delta) + g_{aa}(p_a p_M - \delta)^2}{p_M^2}$$

式中，$G_T$和$G_M$分别为在性状和标记位点的基因型。与其他标记基因型相关的风险用相

似的方式得到。现在除了将在标记位点，而不是疾病位点的频率和风险用于计算外，对标记位点测试的 NCP 的计算步骤完全按以前的方式进行。

## 通过模拟发现关联的计算效能

正如我们所看到的，对遗传关联的统计测试的 NCP 闭式表达在一定情景范围内可以快速和便利地用来计算期待效能。然而这种方式仅对最简单的研究设计有用。实际上，数据组通常涉及更复杂的情况。举例来讲，一个数据组可能由一系列不同类型的系谱结构组成，或个体可能已根据极端的数量性状被确认。网络上没有在这些复杂情况下做效能计算的工具，对于研究者来说得到 NCP 并独立分析做效能计算也不是一件简单的事。在这些情况下效能计算的实施需要数据模拟的帮助。

数据模拟的基本思想是对于给定的遗传模型的关联的替换假设产生成千的重复数据组。换句话说，每一个重复的数据组就像从同一个潜在群体中随机提取一次样品。数据的产生可能涉及模拟一系列不同的进程（举例来讲，从父母向子女传递等位基因）或是根据一些模拟性状值确定个体。然后用每个重复数据组进行遗传关联测试，关联的证据超越了临界值的重复部分作为测试效能的评估。因此模拟对于发现遗传关联的估计效能是一种非常灵活的策略。其缺点是需要大量的计算，对每个特定的群体参数组需要产生成千的重复样品。

## 影响发现关联效能的因素

一系列因素可能影响发现关联的效能。这包括遗传效应的大小、遗传效应的遗传方式、疾病的流行程度、事例与对照的比例、基因频率、标记和性状位点之间连锁不平衡的程度等。在下面所有讨论的情况下，我们用 Purcell 在线遗传效能计算器（genetic power calculator)(http://pngu. mgh. harvard. edu/～purcell/gpc; Purcell et al. 2003）来确定发现关联的期待效能。

### 遗传方式

表 3-3 陈列了不同的遗传模式、基因型相对风险和致病基因频率对发现关联效能的影响。该表明确地显示为了达到 90％的发现关联效能，在显著性阈值为 $\alpha=0.05$ 或 $\alpha=5\times10^{-7}$ 时所需的父母-子女三方（事例/对照组）的数量。$\alpha=0.05$ 的显著水平在一个候选基因研究中分析单一多态时可能是合适的，而 $\alpha=5\times10^{-7}$ 的显著水平在基因组范围的关联研究中将更加合适。在这两种情况下，我们假定一个相对稀少的疾病（$K=0.001$），在性状和标记位点之间完全的连锁不平衡（$D'=1$），标记与疾病位点具有同样的等位频率（这事实上是说我们直接确定基因型并测定疾病本身的位点）。对于事例/对照组，关联的测试是等位基因独立的 Pearson 的卡方检验（Sasieni 1997）。对于父母-子女三方，相关的测试是 Spielman 的 TDT（Spielman et al. 1993）。

**表 3-3 在显著性阈值为 $\alpha=0.05$ 或 $\alpha=5\times10^{-7}$ 时，为使发现关联的效能达到 90%所需的(*A*)个体例子或(*B*)父母-子女三方的数量**

A. 个体例子

| | | $\alpha=0.05$ | | | | $\alpha=5\times10^{-7}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 积乘的 | 相加的 | 隐性的 | 显性的 | 积乘的 | 相加的 | 隐性的 | 显性的 |
| $GRR_{Aa}$ | $p_A$ | | | | | | | | |
| 4.0 | | | | | | | | | |
| | .01 | 301 | 312 | $1.195\times10^{6}$ | 318 | 1137 | 1179 | $4.525\times10^{6}$ | 1201 |
| | .10 | 40 | 53 | 1531 | 64 | 150 | 201 | 5796 | 242 |
| | .50 | 27 | 60 | 55 | 195 | 101 | 225 | 207 | 736 |
| | .80 | 60 | 153 | 71 | 2611 | 225 | 577 | 267 | 9888 |
| 2.0 | | | | | | | | | |
| | .01 | 1602 | 1602 | $1.065\times10^{7}$ | 1656 | 6065 | 1305 | $4.032\times10^{7}$ | 6270 |
| | .10 | 190 | 190 | $1.007\times10^{4}$ | 263 | 718 | 718 | $4.678\times10^{4}$ | 995 |
| | .50 | 92 | 92 | 260 | 512 | 348 | 348 | 983 | 1936 |
| | .80 | 175 | 175 | 233 | 6099 | 661 | 661 | 881 | $2.31\times10^{4}$ |
| 1.5 | | | | | | | | | |
| | .01 | 5319 | 5179 | $3.462\times10^{7}$ | 5465 | $2.014\times10^{4}$ | $1.961\times10^{4}$ | $3.462\times10^{7}$ | $2.069\times10^{4}$ |
| | .10 | 609 | 483 | $3.913\times10^{4}$ | 800 | 2306 | 1826 | $3.913\times10^{4}$ | 3032 |
| | .50 | 260 | 126 | 691 | 1266 | 983 | 477 | 3207 | 4795 |
| | .80 | 457 | 192 | 527 | $1.406\times10^{4}$ | 1728 | 725 | 2447 | $5.325\times10^{4}$ |

B. 父母-子女三方

| | | $\alpha=0.05$ | | | | $\alpha=5\times10^{-7}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 积乘的 | 相加的 | 隐性的 | 显性的 | 积乘的 | 相加的 | 隐性的 | 显性的 |
| $GRR_{Aa}$ | $p_A$ | | | | | | | | |
| 4.0 | | | | | | | | | |
| | .01 | 304 | 315 | $1.197\times10^{6}$ | 321 | 1150 | 1192 | $4.534\times10^{6}$ | 1214 |
| | .10 | 43 | 56 | 1537 | 67 | 160 | 212 | 5819 | 253 |
| | .50 | 30 | 62 | 58 | 198 | 111 | 235 | 217 | 748 |
| | .80 | 63 | 156 | 74 | 2619 | 235 | 588 | 278 | 9917 |
| 2.0 | | | | | | | | | |
| | .01 | 1608 | 1608 | $1.067\times10^{7}$ | 1662 | 6089 | 6089 | $4.04\times10^{7}$ | 6294 |
| | .10 | 193 | 193 | $1.238\times10^{4}$ | 266 | 730 | 730 | $4.688\times10^{4}$ | 1007 |
| | .50 | 95 | 95 | 263 | 515 | 359 | 359 | 995 | 1950 |
| | .80 | 178 | 178 | 236 | 6114 | 672 | 672 | 893 | $2.315\times10^{4}$ |
| 1.5 | | | | | | | | | |
| | .01 | 5333 | 5193 | $3.469\times10^{7}$ | 5479 | $2.02\times10^{4}$ | $1.967\times10^{4}$ | $1.612\times10^{8}$ | $2.075\times10^{4}$ |
| | .10 | 613 | 486 | $4.811\times10^{4}$ | 805 | 2321 | 1840 | $1.822\times10^{5}$ | 3048 |
| | .50 | 263 | 129 | 852 | 1271 | 995 | 488 | 3223 | 4814 |
| | .80 | 460 | 195 | 651 | $1.409\times10^{4}$ | 1741 | 736 | 2462 | $5.337\times10^{4}$ |

很明显，当效应量大时发现关联的效能也达到最大。事例/对照和父母-子女的设计可以较方便地发现具有中等到大的效应的常见等位基因（基因型相对风险>2）。在多数情景中，只需要几十或几百个事例/（家庭）三方；反之，为发现小的效应位点（基因型的相对风险约为 1.5，这对于多数复杂性状或疾病来说或许是可预计的效应量），应该需要更大的样品量。对于基因组范围关联的研究尤为如此，这常需要上千个体的样品量以便具有适当的效能发现小效应等位基因。有意思的是，这样考虑时尽管所要求的显著水平在 5 个数量级水平上变动，所需要的样品量仅仅增加大约 5 倍（需要的样品约与 α 水平的对数相当（Witte et al. 20001)）。通常，当致病基因的频率居中时关联的效能最大，但这在一定程度上依赖于潜在的遗传方式。例如，对于显性疾病，当致病基因频率高时需要更多的个体。发现稀有变异的效能在通常情况下较低。

表 3-3 也显示了对于各种各样的疾病模型，对于给定的效能水平，当疾病稀少时发现关联所需要的（家庭）三方的数量大约等于个体例子的数量（假定事例和对照个体相等）。换句话说，在疾病稀有的情况下，事例/对照研究中需要基因型分型的个体总数大约是三方研究所需的 2/3（Bacanu et al. 2000）。对于所分析的个体和标记数都非常大（因此所需经费数目也大）的基因组范围的关联研究，这一减少得到特别关注。不过必须记住与事例/对照设计相比，基于家庭的关联测试确实有一些优点。例如，TDT 控制了群体的分层效应，该效应在事例/对照设计中可能产生假关联，但仍允许发现起源于父母的效应和基因型的错误。

显示了积乘的、相加的、隐性的和显性的模型的结果。假定事例和对照个体具有相同数目、疾病流行 $K=0.001$。

## 疾病的流行

当疾病流行增加时，运用事例/对照设计发现关联的效能也跟着增加。这是因为在对照个体中高风险基因频率减少（假定对照个体是根据没有风险等位基因的标准选出的），因此在事例和对照组之间的基因频率差别增大。这样，对于常见疾病，假设同样的潜在遗传模式存在，TDT 所需的个体量比事例/对照的关联测试多得多。然而，只要基因型的相关风险和遗传模式保持恒定，疾病的流行不影响 TDT 的效能。这是因为 TDT 使用的基于家庭的对照不被疾病的存在/不存在所监控（父母的遗传组成不会因为疾病的流行而改变）。

## 事例和对照个体的比例

用事例/对照进行关联测试的效能也随着样品中受累和非受累个体的相对比例而变。一般地，对于给定的个体量，最有利的比例是事例和对照相等。然而在一些情况下，确定受累个体可能是困难的或昂贵的。这时补充更多的对照个体仍然可以增加关联测试的效能。McGinnis 等（2002）指出在假设对照量无穷多的情况下，只要有一半的事例就能达到与用等量的事例和对照研究同样的效能。增加 3～5 倍的对照个体量就能接近这一潜在能力。最后，事例和对照的最适比例将取决于获得每种的相对支出（McGinnis et al. 2002）。

## 连锁不平衡和标记基因频率

当连锁不平衡很高时（祖先在标记和性状位点之间很少发生重组），并且当性状和标记位点有相似的等位基因频率时发现关联的效能最大。由于在 dbSNP（Sachidanandam et al. 2001）和 HapMap 数据库（Mtshuler et al. 2005）所发现的多数 SNP，以及由此推出的那些商业上可利用的标记集合都是常见的变异（较少的等位基因频率＞5%），这就使通过遗传关联发现稀有单位基因的意图出现问题。因此利用这些标记资源进行关联研究可能对发现稀有变异的能力有限制（记住，发现稀有疾病等位基因的能力总是非常低的）。这一限制可以通过与疾病基因可能具有相近频率的临近 SNP 单倍型的形成加以说明。这种方式不完全令人满意（Lin et al. 2004）。

## 基因组范围的关联

高通量技术的新进展、基因型分型费用的降低（Matsuzaki et al. 2004），以及国际人类基因组单体型图（Haplotype Map）的出版（Altshuler et al. 2005）使通过在基因组范围扫描数千个体数十万的多态性来寻找复杂疾病基因成为可能。第一批出现在文献中的基因组范围的关联研究结果令人鼓舞（Ozaki et al. 2002；Cheung et al. 2005；Klein et al. 2005；Maraganore et al. 2005）。尽管这些设计很有希望分析复杂性状和疾病的遗传基础，但其中还有一系列统计的问题，其中一个是我们使用两步法检验来减少基因型分型的费用。

为减少对数千对象使用数十万的标记作基因型分型的费用，一个在费用上更有利的策略可能是分阶段对个体作基因型分型。最开始一部分个体（$\pi_{individuals}$）在所有标记（$N_{markers}$）上进行基因型分型。然后再将结果最有希望的一部分标记（$\pi_{markers}$）用来对剩余个体进行分型（Sobell et al. 1993）。与用所有标记对所有个体进行分型的一步方法相比，两步法可以在保留其效能的前提下减少相当的分型费用（Satagopan et al. 2002，2004；Satagopan and Elston 2003；Maraganore et al. 2005；Thomas et al. 2005；Skol et al. 2006）。

Skol 等（2006）最近检验了用于分析从两步基因组范围关联扫描得来的数据的两个不同方法。在“基于复制”（replication-based）策略中，只对在步骤的第二阶段作了基因型分型的标记和个体进行关联测试。为了保证基因组范围的错误率在 $\alpha=0.05$ 之下，运用了 Bonferonni 修正的 $\alpha_{genome}=0.05/(N_{markers}\times\pi_{markers})$ 显著性水平。在“共同分析”（joint analysis）策略中，将第一阶段和第二阶段的测试统计相结合并与 $\alpha_{genome}/N_{markers}$ 的近似显著性水平作比较。研究发现尽管有更严格的显著性水平，与“基于复制”策略相比，对从两个阶段来的数据进行共同分析的方法几乎都能导致发现关联效能的增加。在适当选择的阈值下，用这种方法提供的效能可与单阶段设计媲美。共同分析的效能在第一阶段样品分型的比例下降时下降，推测可能是因为偏向于疾病的变异不太可能在第二阶段被选择出来进行分型。类似地，由于减少了让真实风险变异通过下一阶段的可能性，使用太少的标记通过第二阶段也会导致共同分析的效能降低。相反，在“基于复制”的策略中，当较少的标记用于随后的测试时共同分析能力增加，这是因为由于多重

测试进行了较少的统计测试而因此受到较少的处罚。事实上，只在一种情况下共同分析不如“基于复制”的策略有力，即当第二阶段关联的强度远强于第一阶段时。因此推荐两阶段全基因组关联扫描由共同分析策略进行分析。此时大部分样品在第一阶段作了基因型分型（$\pi_{individuals}>30\%$），而相对大部分的标记被选用在第二阶段作后续分析（$\pi_{markers}>1\%$）。统计遗传学中心（the Center for Statistical Genetics）提供了有用的网络工具（http://csg. sph. umich. edu），它可被研究者用来计算在两步法设计中发现关联的效能。

## 结论

我们总结出一些基本的指导，研究者可能在进行遗传关联研究的设计或进行分析阶段效能计算时发现它们有用。

（1）在研究的设计阶段，使用诸如 Purcell's GPC（http://pngu. mgh. harvard. edu/～purcell/gpc/）之类的网络工具计算在疾病已知（举例来说，遗传方式、效应的大小）的情况下将要作基因型分型的期待个体数。如果一个人正在进行候选基因研究并愿意假设一个功能变异位于正在分型的区域的某处（举例来说，一个过去的连锁研究可能提示一个疾病位点位于这一区域），那么可以有根据地假设在标记和疾病位点之间存在高水平的连锁不平衡。相反，如果正在计划一个基因组范围的关联研究，那么就必须考虑到一个起因变异可能与任何标记都没有可感知的连锁不平衡。商业上用可用的标记集合捕获常见变异的效能已经被 Barrett 和 Cardon（2006）量化。

（2）如果正从事复杂的研究设计，可能需要实施数据模拟，以计算期待的效能。

（3）如果正在考虑一个全基因组关联研究，两阶段设计是一种有效和经济的对对象进行基因型分型的方法，它可以保持发现关联的效能。通常大量的样品应该在第一阶段作基因型分型（>30%），相对大量的标记应该被选出用于跟随研究（>1%），数据应该用共同分析策略进行分析。一个在两阶段设计中可被用来计算发现关联效能的网络工具位于 http://csg. sph. umich. edu。

## 参考文献


Altshuler D., Brooks L.D., Chakravarti A., Collins F.S., Daly M.J., and Donnelly P. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Bacanu S.A., Devlin B., and Roeder K. 2000. The power of genomic control. *Am. J. Hum. Genet.* **66:** 1933–1944.

Barrett J.C. and Cardon L.R. 2006. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38:** 659–662.

Cheung V.G., Spielman R.S., Ewens K.G., Weber T.M., Morley M., and Burdick J.T. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437:** 1365–1369.

Devlin B. and Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29:** 311–322.

Gauderman W.J. 2002. Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidemiol.* **155:** 478–484.

Gordon D., Finch S.J., Nothnagel M., and Ott J. 2002. Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. *Hum. Hered.* **54:** 22–33.

Klein R.J., Zeiss C., Chew E.Y., Tsai J.Y., Sackler R.S., Haynes C., Henning A.K., SanGiovanni J.P., Mane S.M., Mayne S.T., et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308:** 385–389.

Lewontin R.C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49:** 49–67.

Lin S., Chakravarti A., and Cutler D.J. 2004. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.* **36:** 1181–1188.

Maraganore D.M., de Andrade M., Lesnick T.G., Strain K.J., Farrer M.J., Rocca W.A., Pant P.V., Frazer K.A., Cox D.R., and Ballinger D.G. 2005. High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.* **77:** 685–693.

Matsuzaki H., Loi H., Dong S., Tsai Y.Y., Fang J., Law J., Di X., Liu W.M., Yang G., Liu G., et al. 2004. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14:** 414–425.

McGinnis R., Shifman S., and Darvasi A. 2002. Power and efficiency of the TDT and case-control design for association scans. *Behav. Genet.* **32:** 135–144.

Mitra S.K. 1958. On the limiting power function of the frequency chi-square test. *Ann. Math. Stat.* **29:** 1221–1233.

Ozaki K., Ohnishi Y., Iida A., Sekine A., Yamada R., Tsunoda T., Sato H., Sato H., Hori M., Nakamura Y., and Tanaka T. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32:** 650–654.

Purcell S., Cherny S.S., and Sham P.C. 2003. Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19:** 149–150.

Sachidanandam R., Weissman D., Schmidt S.C., Kakol J.M., Stein L.D., Marth G., Sherry S., Mullikin J.C., Mortimore B.J., Willey D.L., et al. (International SNP Map Working Group). 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.

Sasieni P.D. 1997. From genotypes to genes: Doubling the sample size. *Biometrics* **53:** 1253–1261.

Satagopan J.M. and Elston R.C. 2003. Optimal two-stage genotyping in population-based association studies. *Genet. Epidemiol.* **25:** 149–157.

Satagopan J.M., Venkatraman E.S., and Begg C.B. 2004. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* **60:** 589–597.

Satagopan J.M., Verbel D.A., Venkatraman E.S., Offit K.E., and Begg C.B. 2002. Two-stage designs for gene-disease association studies. *Biometrics* **58:** 163–170.

Schork N.J. 2002. Power calculations for genetic association studies using estimated probability distributions. *Am. J. Hum. Genet.* **70:** 1480–1489.

Sham P.C., Cherny S.S., Purcell S., and Hewitt J.K. 2000. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **66:** 1616–1630.

Skol A.D., Scott L.J., Abecasis G.R., and Boehnke M. 2006. Corrigendum: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38:** 390.

Sobell J.L., Heston L.L., and Sommer S.S. 1993. Novel association approach for determining the genetic predisposition to schizophrenia: Case-control resource and testing of a candidate gene. *Am. J. Med. Genet.* **48:** 28–35.

Spielman R.S., McGinnis R.E., and Ewens W.J. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52:** 506–516.

Thomas D.C., Haile R.W., and Duggan D. 2005. Recent developments in genomewide association scans: A workshop summary and review. *Am. J. Hum. Genet.* **77:** 337–345.

Witte J.S., Elston R.C., and Cardon L.R. 2000. On the relative sample size required for multiple comparisons. *Stat. Med.* **19:** 369–372.

# 互联网资源

**http://csg.sph.umich.edu** The CaTS Power Calculator software at the Center for Statistical Genetics Web site is a simple, multi-platform interface for carrying out power calculations for large genetic association studies, including two-stage genome-wide association studies.

**http://hydra.usc.edu/GxE** Gene × Environment, Gene × Gene Interaction Home page

**http://linkage.rockefeller.edu/pawe** The Power for Association With Error (PAWE) program is designed to perform asymptotic power and sample size calculations for genetic case/control studies with a diallelic locus (for example, a SNP) in the presence of errors.

**http://pngu.mgh.harvard.edu/~purcell/gpc/** The on-line Genetic Power Calculator provides automated power analysis for variance components, quantitative trait loci linkage and association tests in sibships, and other common tests.

# 4　遗传分析：在连锁和关联之间

Albert Vernon Smith

*Genthor ehf., 101 Reykjavik, Iceland, and Icelandic Heart Association, 201 Kopavogur, Iceland*



## 简介

在人类疾病遗传学研究中，用于确定与疾病相关的基因或基因组区域的方法可大致分为两类：连锁分析和遗传关联分析。以家庭为基础的连锁分析研究是成功地将许多孟德尔遗传病进行基因作图的基础，但这种方法在寻找中度危险性的等位基因时显得能力不足，而这类等位基因很可能与复杂疾病相关。连锁分析特别适合于具有明显家族特征和高度外显率的疾病。在连锁分析时，研究者从有许多成员染病的系谱中鉴定和收集个体资料，然后研究者分析一套覆盖整个基因组和系谱的标记，寻找那些受累者共有的、有较高频率的染色体区域（Ott 1999）。用这种方法已经确定了大量的疾病位点，特别是具有孟德尔遗传特点的那些性状。虽然这些方法获得了显著的成功，但所研究的疾病多是罕见的，不能完全反映医学实践中的大量疾病。连锁研究在寻找微效多基因效应时作用有限，而这种效应被认为和一些常见疾病，如心脏病、哮喘和糖尿病等有关。

寻找致病基因的另一种替代方法是遗传关联研究。在进行这种分析时，先收集具有某特殊性状的大量个体，寻找那些比对照更常见的个体标记等位基因或基因型（Cardon and Bell 2001）。可以根据所测试的位点和在整个群体中与疾病有关的变异之间的连锁不平衡（LD）在无关的个体中进行关联分析。与对照相比，这些与疾病有关的等

位基因在患者中更常见（或者说与他们处于连锁不平衡中）。在最近能够同时进行成百上千个标记的基因型分型的基因型系统问世之前，研究者已经对一小部分候选基因进行了典型的关联分析，这些候选基因被认为与关系到特定疾病发生的生物学进程相关。多年来科学家在许多研究中已经进行过这种尝试，但大量的独立研究只确定了少量的基因与疾病有确切关联。

虽然可以找出这两种方法之间的区别，但它们通常是交错的。典型的情况是，一个连锁研究可能最开始确定了一个与特定疾病连锁的位点，但遗传连锁分析的局限之一是成功率很低，并且只能确定与疾病相关的覆盖了数百万碱基的 DNA 区域。一旦与疾病相关的区域被连锁分析确定下来，接下来一般就要通过遗传关联来处理这一信息，这是因为遗传关联具有更细致的分析力。在确定了疾病相关位点之后，在感兴趣区域的一系列候选基因将被挑出来，然后对此位点内的标记作对照组和疾病组的基因型分析，寻找在疾病组中更常出现的特异性标记等位基因。

HapMap 产品是人类疾病遗传学研究中的革命性进展，特别是对所谓常见疾病的研究（International HapMap Consortium 2005）。HapMap 利用密集的遗传作图，并发展可用于商业的高致密的基因型产品，在遗传关联研究领域中促进了范式的转移。现在使用有 500 000～1 000 000 个单核苷酸多态的基因型的产品可以对全基因组，而不仅仅是对有限的候选基因进行遗传关联研究。

尽管近来全基因组关联分析已成为可能，并在常见疾病的研究中比连锁分析更具优势，但熟悉两者仍是重要的。基于家庭的材料可用来改善关联分析，而遗传关联仍将继续作为连锁分析最初确定的信息，在后续研究中使用。本章试图概括介绍进行遗传连锁和遗传关联研究所需的工具。

## 为遗传分析创造一个输入文件

虽然没有为遗传分析包预备的单一标准的输入文件格式，但是用于 LINKAGE 程序（Lathrop et al. 1984）的文件格式使用相对较多，可以用来输入下述不同的遗传分析包。每一种格式有一些小的不同，但它们的基本设计一致。

### 描述关系

输入遗传分析资料时必须描述个体之间的关系。尽管系谱通常复杂，但是基本的信息可以减少到 5 项：一个家庭标识符、一个个体标识符、每个父母的标识符（如果知道）和个体的性别。因此需要建立一个系谱文件以描述这类信息。例如，在一个有一对同胞、他们的父母和父亲的父母的小系谱中，基本的信息可以描述如下。

| 家庭 | 个人 | 父亲 | 母亲 | 性别 |
|---|---|---|---|---|
| example | grandpa | unknown | unknown | m |
| example | grandma | unknown | unknown | f |
| example | father | grandpa | grandma | m |
| example | mother | unknown | unknown | f |

| | | | | |
|---|---|---|---|---|
| example | sister | father | mother | f |
| example | brother | father | mother | m |

（上述中，grandpa：祖父；grandma：祖母；father：父亲；mother：母亲；sister：姐妹；brother：兄弟；unknown：未知；m：男性；f：女性）

实践时这些文件标识符由特定的数值替代。在使用了特定整数并用 1（男性）或 2（女性）重新编码的性别符号后，上述例子变成如下情况。

| | | | | |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 |
| 1 | 2 | 0 | 0 | 2 |
| 1 | 3 | 1 | 2 | 1 |
| 1 | 4 | 0 | 0 | 2 |
| 1 | 5 | 3 | 4 | 2 |
| 1 | 6 | 3 | 4 | 1 |

## 描述表型和基因型

通常这 5 个标准项之后是表型和标记基因型。加入一个框来包含表型信息。对于不连续的性状，习惯上用 1 表示未受累、2 表示受累、0 表示没有信息。对于数量性状，插入实际数值。然后加入标记基因型，用两个连续整数表示，每个对应一个等位基因。对于两个等位的 SNP，可用 1 表示其中一个、2 表示另一个、0 表示没有信息。下述例子中，加入了疾病描述、数量性状和两个标记的信息。

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 1 | x | 1 | 2 | 2 | 1 |
| 1 | 2 | 0 | 0 | 2 | 1 | x | 1 | 1 | 1 | 1 |
| 1 | 3 | 1 | 2 | 1 | 1 | x | 1 | 1 | 2 | 1 |
| 1 | 4 | 0 | 0 | 2 | 1 | x | 2 | 2 | 1 | 1 |
| 1 | 5 | 3 | 4 | 2 | 2 | 5.6 | 1 | 1 | 2 | 1 |
| 1 | 6 | 3 | 4 | 1 | 2 | 1.3 | 2 | 1 | 2 | 1 |

通常系谱文件用带有 a. *ped* 的文件扩展名保存。

## 描述系谱文件

由于上述文件包含了疾病状态、数量性状变异的特定数字和基因型标记等，必须构建一个伴随数据文件以描述系谱文件的内容。尽管对不同的程序在使用细节上略有不同，对于系谱文件的每个数据项，数据文件将有一行，表示数据类型（M：标记；A：疾病状态；T：数量性状；C：共变量）和每一项的标记。上述系谱文件有一个疾病状态、一个数量性状和两个标记基因型，将构建下述数据文件。

A some disease（某种疾病）

T some trait（某种性状）

M marker one（标记一）

M marker two（标记二）

通常数据文件用带有 a. *dat* 的文件扩展名保存。在这个 *dat* 文件中应含有与特定的

*ped* 文件相应的行数。

### 遗传作图信息

对于许多遗传分析来说，知道标记的相对位置很重要。对于连锁分析程序来讲，需要提供遗传图位置，因为这是分析时用到的因素之一。然而对于遗传关联分析来讲，常提供物理位置，而不提供遗传图位置。连锁分析依赖遗传位置作为分析的一部分，而许多遗传关联测试独立地检测每一个标记，因此不需要遗传图的位置。典型的作图文件列出了染色体、标记名称、遗传位置和（随意的）物理位置。

| 染色体 | 标记 | 图 | 物理 |
|---|---|---|---|
| 24 | 标记一 | 14.1 | 12 000 000 |
| 24 | 标记二 | 15.3 | 13 500 000 |

通常这个作图文件用带有 a. *map* 的文件扩展名保存。在一些情况下，作图文件必须有与 *ped* 文件中标记数相一致的行数；而在另一些情况下，非典型的标记可以忽略，作图位置只在标记也列在 *dat* 文件时才使用。

### 用这些文件工作

虽然不同的遗传分析程序使用的文件格式略有不同，基本要素十分相似，通常由下列文件组成：一个带有系谱信息的 *ped* 文件（包括家庭关系、疾病状态和标记基因型的信息）、一个描述系谱文件内容的 *dat* 文件和一个给出了针对被分型的标记信息的作图文件。对一个给定的分析，这三个文件形成一组。对于不同的遗传分析包，所需要的确切的框可能稍有不同，但都使用这一基本的文件结构，包括下面所描述的。

## 用 MERLIN 进行连锁分析

MERLIN 是一个免费的应用程序，能够进行系谱数据的分析（Abecasis et al. 2002）。与许多其他用作系谱数据分析的包不同，MERLIN 使用的运算法则特别适合于致密的遗传作图。MERLIN 是一个完整的包，可以通过许多方式分析系谱资料，包括对血缘一致性（identity-by-descent，IBD）和血族关系系数的计算，实施非参数和变异成分连锁分析，发现错误，对信息内容作图等。下面描述如何用 MERLIN 进行连锁分析，以便初步了解这个软件包。连锁分析检测染色体区域和感兴趣的性状的共分离，按如下步骤进行基本的连锁分析。

（1）从 http://www. sph. umich. edu/csg/abecasis/MERLIN/下载和安装 MERLIN 软件包。Linux、Windows、Sun 和 Mac 可以使用二进制，而且可以利用源编码，必要时可以编辑 MERLIN。软件包中的程序可从命令行启动。

（2）创建 3 个适当的输入文件：一个数据文件（*file. dat*）、一个系谱文件（*file. ped*）和一个作图文件（*file. map*）。输入文件如上所述，虽然在作图文件中没有包含物理位置（这个文件中需要 3 个框：染色体、标记名称和遗传位置）。

（3）一旦创立了这些文件，为确保文件的正确运行需进行测试。为此运行 MER-

LIN 中的 pedstats 程序。它需要输入一个 *dat* 文件（-d 参数）和 *ped* 文件（-p 参数）。运行如下：

```
pedstats -d file. dat -p file. ped
```

如果这些文件格式适当，研究者将得到这些文件的概要统计。在分配的位置提供样品文件，运行这些测试文件可得到从这个程序输出的很好的样品（MERLIN 的其他程序也一样）。

（4）如果 pedstats 运行正常，用如下命令行参数运行 MERLIN：an input *dat* file (-d)、a *ped* file (-p) 和 a *map* file (-m)，然后指定分析的类型。MERLIN 具有一系列的选项，但通常最好从非参数连锁（NPL）分析开始，NPL 可由--npl 命令行选项指定。MERLIN 进行的标准 NPL 分析使用 Kong and Cox（1997）线性模型来估计连锁证据。这将作如下运行。

```
merlin -p file. ped -d file. dat -m file. map -npl
```

从使用的选项的概要开始输出：

```
MERLIN 1.0.1 - (c) 2000-2005 Goncalo Abecasis

 References for this version of Merlin:

    Abecasis et al (2002) Nat Gen 30:97-101         [original citation]
    Fingerlin et al (2004) AJHG 74:432-43           [case selection for association studies]
    Abecasis and Wigginton (2005) AJHG 77:754-67    [1d modeling, parametric analyses]

 The following parameters are in effect:
                     Data File :         file.dat (-dname)
                 Pedigree File :         file.ped (-pname)
            Missing Value Code :          -99.999 (-xname)
                      Map File :         file.map (-mname)
            Allele Frequencies : ALL INDIVIDUALS (-f[a|e|f|m|file])
                   Random Seed :           123456 (-r9999)

 Data Analysis Options
                       General : --error, --information, --likelihood, --model [param.tbl]
                    IBD States : --ibd, --kinship, --matrices, --extended, --select
                   NPL Linkage : --npl [ON], --pairs, --qtl, --deviates, --exp
                    VC Linkage : --vc, --useCovariates, --ascertainment
                   Haplotyping : --best, --sample, --all, --founders, --horizontal
                 Recombination : --zero, --one, --two, --three, --singlepoint
                     Positions : --steps, --maxStep, --minStep, --grid, --start, --stop
               Marker Clusters : --clusters [], --distance, --rsq, --cfreq
                        Limits : --bits [24], --megabytes, --minutes
                   Performance : --trim, --noCoupleBits, --swap, --cache []
                        Output : --quiet, --markerNames, --frequencies, --perFamily, --pdf,
                                 --prefix [merlin]
                    Simulation : --simulate, --reruns, --save
```

然后出现分析结果。

```
Phenotype:affection [All] (200 families)
==================================================
          Pos      Zmean     pvalue      delta       LOD     pvalue
          min     -20.00        1.0     -0.707    -60.21        1.0
          max      20.00    0.00000      0.707     60.21    0.00000
        0.000       0.96        0.2      0.092      0.27       0.13
```

```
 5.268    1.39    0.08     0.126   0.54   0.06
10.536    1.27    0.10     0.110   0.43   0.08
15.804    1.43    0.08     0.128   0.56   0.05
21.072    0.88     0.2     0.083   0.22    0.2
26.340    1.37    0.08     0.130   0.55   0.06
31.608    1.53    0.06     0.151   0.71   0.04
36.876    2.18   0.014     0.197   1.32  0.007
42.144    2.60   0.005     0.218   1.75  0.002
47.412    3.00  0.0014     0.251   2.33 0.0005
52.680    3.43  0.0003     0.286   3.05 0.00009
...
```

在这个输出中，前两行指出这组数据的最大可能得分，下面各行是每一个标记位置的结果。在这个例子中，在 52.680 出现连锁峰，其 Z 计分是 3.43（0.0003 的渐进 p-值），对应于 Kong and Cox LOD3.05 的计分和 0.00 009 的概率，LOD 计分是遗传连锁对比没有连锁区域的拟然率的对数值。

列在此处的方案只是对十分复杂的软件包的基本介绍，该软件包可以进行系谱数据的多重分析。对每一个选项的详细描述超出了本章的范围。关于如何进行进一步分析的细节，可参考在线文件的详细解答。以上输出罗列了所有可利用的选项，读者从中可以体会到多重的可能性。

应特别注意的是，MERLIN 能够分析模拟的染色体资料，这取决于输入的家庭结构和标记间隔。这类分析可以预计有多少高度相似的峰存在，这取决于所检查的基因型和可利用的标记图。此外，MERLIN 在察觉基因型错误上具有一系列的特点。错误察觉系统通过排除有问题的基因型，可显著改善连锁分析的能力。通过系谱的标记传递信息也可以重建单倍型。IBD 分析可用于估计系谱中任意两个个体从建立者遗传了同一染色体的可能性，包含在 MERLIN 中的许多功能可显著增强连锁分析能力，但在此未做详细描述。

## 用 PLINK 进行关联分析

PLINK 是最近特别为全基因组关联分析研究而设计发展的免费软件包（Purcell et al. 2007）。该软件可以进行多种对解读基因组范围的关联研究而言非常关键的分析，包括质量和数量性状的分析。它的一些特点可帮助数据处理和质量控制，并可采用众多方法进行关联分析。

使用 PLINK 进行基本的关联分析。

（1）从 http://pngu. mgh. harvard. edu/purcell/plink/下载并安装 PLINK，Linux、Windows、Sun 和 Mac OS X 可以使用二进制，而且可以利用源编码，必要时可以编辑 PLINK。

（2）为 PLINK 创建一系列合适的输入文件。如上所述构建一个 *ped* 文件。PLINK

只需要一个疾病状态的框，位于含有基因型的框之前（对于 PLINK，默认的缺失表型值是－9）。构建一个 *map* 文件。对 *ped* 文件中的每一个标记 PLINK 需要对应一行。*map* 文件含有 4 个框：染色体、标记名、遗传位置和物理位置。在许多分析中用不到遗传位置框，通常可以在 *map* 文件中设为 0。

（3）为运行 PLINK 并检测输入文件的完整性，使用如下的命令，这将得到一个简要的统计量。

plink --file hapmap1

（所列的输入文件来自一个在线 PLINK 指南，包含了一个模拟的疾病变量 rs2222162）

在这个例子中，-- file 选项表示 *ped* 文件 hapmapl. ped 和 *map* 文件 hapmap 1. map。每次运行 PLINK 时有一个详细的日志文件的输出，这给予运行所需的选项。就上面所给的例子来讲，可观察到如下输出。

```
@----------------------------------------------------------@
|        PLINK!        |     v0.99q     |   17/Jan/2007     |
|----------------------------------------------------------|
|  (C) 2007 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|       http://pngu.mgh.harvard.edu/purcell/plink/          |
@----------------------------------------------------------@

Web-based version check ( --noweb to skip )
Connecting to web...  OK, v0.99q is current

*** Pre-Release Version ***

Writing this text to log file [ plink.log ]
Analysis started: Mon Apr 23 12:15:31 2007

Options in effect:
        --file hapmap1

83534 (of 83534) markers to be included from [ hapmap1.map ]
89 individuals read from [ hapmap1.ped ]
89 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
44 cases, 45 controls and 0 missing
89 males, 0 females, and 0 of unspecified sex
Before frequency and genotyping pruning, there are 83534 SNPs
Applying filters (SNP-major mode)
89 founders and 0 non-founders found
0 of 89 individuals removed for low genotyping ( MIND > 0.1 )
Total genotyping rate in remaining individuals is 0.99441
859 SNPs failed missingness test ( GENO > 0.1 )
16994 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 65803 SNPs
```

## 使用 PLINK 时的其他考虑

（1）因为全基因组相关数据通常具有大量的标记，建立比开启 *ped*/*map* 文件更紧凑的数据二进制表示非常有用。对后来的分析来说二进制表示装载得也更快。为做到这一点可用如下命令。

```
plink --file hapmap1 --make-bed --out hapmap1
```

为用这些文件工作，用--bfile 选项替换--file 命令行选项。对所有的命令使用--out 选项来指定植根于输出的文件名。

(2) 使用命令行选项可以产生一系列不同的概要统计表，包括缺失率（--missing flag）和基因频率（--freq）的信息。

(3) 通过添加--assoc 命令行标记（command line flag）进行基本的相关分析。基本的命令是：

```
plink --bfile hapmap1 --assoc --out as1
```

这产生一个输出文件 as1. assoc 并开始如下输出。

```
CHR        SNP  A1      F_A       F_U  A2     CHISQ        P       OR
  1  rs6681049   1   0.1591    0.2667   2     3.067  0.07991   0.5203
  1  rs4074137   1  0.07955   0.07778   2  0.001919   0.9651    1.025
  1  rs1891905   1   0.4091       0.4   2   0.01527   0.9017    1.038
...
```

这次输出框里是

- CHR: Chromosome
- SNP: SNP identifier
- A1: Code for allele 1 (the minor, rare allele based on the entire sample frequencies)
- F_A: The frequency of this variant in cases
- F_U: The frequency of this variant in controls
- A2: Code for the other allele
- CHISQ: The chi-squared statistic for this test (1 df)
- P: The asymptotic significance value for this test
- OR: The odds ratio for this test

在 Unix/Linux 环境中，简单使用可利用的命令行工具对相关性统计值列表分类并打印出前 10 个。例如，

```
sort -key = 7 -nr as1. assoc | head
```

这给出

```
13    rs9585021   1    0.625    0.2841   2   20.62   5.586e-06      4.2
 2    rs2222162   1   0.2841    0.6222   2   20.51   5.918e-06   0.2409
 9   rs10810856   1   0.2955   0.04444   2   20.01   7.723e-06    9.016
 2    rs4675607   1   0.1628    0.4778   2   19.93    8.05e-06   0.2125
```

除了提供如上所述的不连续性状分析方法之外，PLINK 能够分析数量性状的相关性。除了直接从 *ped* 文件阅读表型外，PLINK 也可以从一个外部文件阅读表型。这个文件必须指定 3 个框，前两个是在 *ped* 文件中使用的家庭和个人的标识符（个体不一定与 *ped* 文件的顺序相同），第三个框含有表型值，在这个例子中是数量性状的度量。因为 PLINK 自动察觉被检测的性状是不连续的还是数量的，所以运行数量相关分析时使用与上述非常相似的命令行选项。然后可以运行分析。

```
plink --bfile hapmap1 --assoc --pheno qt. phe --out quant1
```

该分析将产生如下输出（在有 *.qassoc* 后缀的文件中）。

| R | SNP | NMISS | BETA | SE | R2 | T | P |
|---|---|---|---|---|---|---|---|
| 1 | rs6681049 | 89 | −0.2266 | 0.3626 | 0.004469 | −0.6249 | 0.5336 |
| 1 | rs4074137 | 89 | −0.2949 | 0.6005 | 0.002765 | −0.4911 | 0.6246 |
| 1 | rs1891905 | 89 | −0.1053 | 0.3165 | 0.001272 | −0.3328 | 0.7401 |
| 1 | rs9729550 | 89 | 0.5402 | 0.4616 | 0.0155 | 1.17 | 0.2451 |

这一文件中的这一区域表示

- 染色体。
- SNP 标识符。
- 这次分析非缺失个体数。
- 回归系数。
- 该系数的标准误。
- 回归 r 平方（多重相关系数）。
- 表型对等位基因数回归的 t-统计值。
- 系数的渐进显著值。

该操作仅仅描述了相关分析这一非常复杂主题的基本要素，应该还有许多其他要素作为分析基因组范围的一部分相关资料而开发出来。PLINK 对相关分析的许多变化具有一些选择，还有一系列的测试以适合群体结构。一些选项可以检测数据和样品质量，以及寻找额外的关系。

已经发展出一个称为 gPLINK 的图形用户界面（GUI），可为一般的 PLINK 操作提供简单的界面。可以使用菜单和对话创建有效的 PLINK 命令，保存命令运行记录，以及跟踪输入输出文件。该界面还在快速发展中，可在 PLINK 主页上找到它的操作细节。对于如何在全基因组相关数据分析中使用 PLINK 可参考在线文件的详细解答。

## 讨论和结论

虽然上面介绍的两个遗传分析包 MERLIN 和 PLINK 最初是为遗传分析的不同方面而设计的，但是它们有许多共同特点。两者都是有多方面用途的软件包，包括错误察觉、概要统计和遗传分析。两者也都得到充分的证明并开放源码程序，这有利于使用者将它们整合进遗传分析工作流程中。

直到最近，揭示疾病遗传因素的研究通常以连锁分析开始。特异性遗传标记的密度不足以在全基因组水平展开相关性扫描。随着 HapMap 以及一系列具有非常致密标记的基因型分析平台的出现，现在已经有可能进行全基因组的相关扫描，寻找与疾病相关的标记。目前使用这一方法开展的研究工作才开始刊登出来。在糖尿病和前列腺癌研究中取得的惊人成功（Gudmundsson et al. 2007；Sladek et al. 2007；Yeager et al. 2007）提示全基因组相关研究将会是一种重要的鉴定新基因的方式，这些基因中的每一个都与疾病关联，具有中等作用。

人们对用全基因组相关分析揭示复杂疾病的遗传因素取得成功寄予希望，然而许多因素使得分析变得复杂。在设计和实施一个关联研究时它们都应该被考虑进去，特别是群体的分层特点。由于发病率和等位基因频率在不同群体中不一致，在一个混合群体中进行遗传相关测试可导致假阳性结果。已经提出许多方案来控制这一因素，只要有高密度的遗传标记，这些方案就会起作用。对于群体结构，基因组控制假定卡方检验的结果被一个常数因子所夸大，并且所有测试可由这个因子进行分组（Devlin and Roeder 1999）。另一个方案涉及结构相关性，个体根据基因型分组，关联的结果由所在分组限定（Pritchard et al. 2000a，b）。还有一个解决方案是进行基于家庭的关联测试，如传递不平衡测试（TDT），该测试不受群体结构影响（Spielman et al. 1993）。

基于家庭的设计可无视群体的混杂而开展，它们对比事例/对照设计还有其他的优点。它们允许连锁和关联测试一起进行，并有利于多重假设检验（就有关的回顾，参见 Laird and Lange 2006）。虽然有这些明显的优点，一个潜在的缺点是收集基于家庭的样品可能较困难。现已发展出一种界面友好、内容全面的基于家庭的关联测试包（family-based association test，FBAT），允许使用者运用家庭内的对照进行疾病表型和单倍型之间的关联/连锁测试（Laird et al. 2000）。这组测试广泛地被调整用于混合群体。这个软件得到充分的证明并可在 http://biosun1. harvard. edu/～fbat/fbat. htm 上免费使用。

虽然连锁和关联研究依赖于不同的设计和方法揭示遗传性疾病，但它们在使用时通常互补。由于连锁研究结果粗糙，连锁确定的位点通常由位点内基因的关联分析加以精确化并确定与所研究的性状相关的基因。虽然全基因组相关分析可用于没有关联的事例/对照样品，这种分析与基于家庭的材料结合考虑可为分析和结果的解释带来好处。因此在可预见的将来，人类疾病的遗传分析将继续依赖连锁分析的各个方面以及关联分析。

## 参考文献


Abecasis G.R. and Wigginton J.E. 2005. Handling marker-marker linkage disequilibrium: Pedigree analysis with clustered markers. *Am. J. Hum. Genet.* **77:** 754–767.

Abecasis G.R., Cherny S.S., Cookson W.O., and Cardon L.R. 2002. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30:** 97–101.

Cardon L.R. and Bell J.I. 2001. Association study designs for complex diseases. *Nat. Rev. Genet.* **2:** 91–99.

Devlin B. and Roeder K. 1999. Genomic control for association studies. *Biometrics* **55:** 997–1004.

Fingerlin T.E., Boehnke M., and Abecasis G.R. 2004. Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. *Am. J. Hum. Genet.* **74:** 432–443.

Gudmundsson J., Sulem P., Manolescu A., Amundadottir L.T., Gudbjartsson D., Helgason A., Rafnar T., Bergthorsson J.T., Agnarsson B.A., Baker A., et al. 2007. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39:** 631–637.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Kong A. and Cox N.J. 1997. Allele-sharing models: LOD scores and accurate linkage tests. Am *J. Hum. Genet.* **61:** 1179–1188.

Laird N.M. and Lange C. 2006. Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* **7:** 385–394.

Laird N., Horvath S., and Xu X. 2000. Implementing a unified approach to family based tests of association. *Genet. Epidemiol.* (suppl. 1) **19:** S36–S42.

Lathrop G.M., Lalouel J.M., Julier C., and Ott J. 1984. Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci.* **81:** 3443–3446.

Ott J. 1999. *Analysis of human genetic linkage*, 3rd edition. Johns Hopkins University Press, Baltimore, Maryland.

Pritchard J.K., Stephens M., and Donnelly P. 2000a. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

Pritchard J.K., Stephens M., Rosenberg N.A., and Donnelly P. 2000b. Association mapping in structured populations. *Am. J. Hum. Genet.* **67:** 170–181.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., de Bakker P.I.W., Daly M.J., and Sham P.C. 2007. PLINK: A toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* (in press).

Sladek R., Rocheleau G., Rung J., Dina C., Shen L., Serre D., Boutin P., Vincent D., Belisle A., Hadjadj S., et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445:** 881–885.

Spielman R.S., McGinnis R.E., and Ewens W.J. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52:** 506–516.

Yeager M. Orr N., Hayes R.B., Jacobs K.B., Kraft P., Wacholder S., Minichiello M.J., Fearnhead P., Yu K., Chatterjee N., et al. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39:** 645–649.

# 互联网资源

**http://biosun1.harvard.edu/~fbat/fbat.htm.** FBAT software implements a broad class of Family Based Association Tests, adjusted for population admixture.

**http://pngu.mgh.harvard.edu/purcell/plink-PLINK**, a free, open-source whole-genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner (see Purcell et al. 2007).

**http://www.sph.umich.edu/csg/abecasis/MERLIN**. MERLIN, a program for analyses of pedigree data. Abecasis Laboratory, Center for Statistical Genetics, University of Michigan.

# 5 NCBI dbSNP 数据库：内容和检索

Michael L. Feolo and Stephen T. Sherry

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894-3804*



## 简介

基因组序列变异与可遗传表型的联系或连锁是在世界范围内得到大量资助的焦点问题。国际人类基因组单体型图（HapMap）计划由科学家们和来自加拿大、中国、日

本、尼日利亚、英国和美国的资助机构的合作进行，以发展公共资源、帮助研究者发现与人类疾病相关的基因从而进行医药品的研制（International HapMap Consortium 2005）。西雅图 SNP 作为国家心脏、肺和血液研究所（NHLBI）基因组应用项目（PGA）的一部分而得到资助。它的重点是为候选基因内的 SNP 和人类潜在刺激反应通路的联系进行鉴定、基因型分型和建模（http://pga.gs.washington.edu）。华盛顿大学国家环境健康科学研究所（NIEHS）的 SNP 项目目标是，作为环境基因组计划的一部分，对环境响应基因的 SNP 进行系统的鉴定和基因型分型（EGP；http://egp.gs.washington.edu）。

在数据处理和统计学推论上，基因组序列变异和可遗传表型的联系也是研究的热点课题（Cardon and Bell 2001；Carlson et al. 2004；Hirschhorn and Daly 2005）。SNP 是遗传变异中最常见的类型之一。当人类某一染色体的任意两个拷贝相互比较时，500～1000 bp发生一次 SNP。生物技术信息国家中心（the National Center for Biotechnology Information，NCBI）的 dbSNP 数据库（Sherry et al. 2001；www.ncbi.nlm.nih.gov/SNP）是世界上最大的变异资料库，数据库根据如下的类型和百分构成对核苷酸序列变异进行分类：①单一核苷酸替代，99.77%；②小的插入/缺失多态性，0.21%；③序列的不变区，0.02%；④微卫星重复，0.001%；⑤指定的变异，<0.001%；⑥未确定的杂合指标，<0.001%。

数据库没有对最小基因频率或多态性功能上的中性提出要求和假定，也没有要求确认状态。这样 dbSNP 的范围包括造成疾病的临床突变和中性多态性等。除了由提交者和 NCBI 确定的记录标识外，dbSNP 条目记录了多态性序列信息，可能给出所观察的多态性的实例、进行实验所需的特定实验条件、包含这一变异的群体的描述和群体或个体基因型的频率信息。本章，你可以用任何种类的核苷酸序列变异替换术语 SNP。

目前普通序列变异的发现提示，选择效应未知的、相对常见的（最小基因频率>5%）SNP 标记在提交报告中占大多数。提交给 dbSNP 的基因型提供了确认、频率和连锁不平衡的信息，以用于商业检测中 SNP 的选择，这些检测为大规模基因组范围内的相关研究所需，也为需要为候选区域（如在第 19 章和第 20 章所讨论的）摘取一套标签 SNP 的较小项目所使用。尽管本章给予的例子通常涉及人类资料，这些资料的结构和概念在非人类物种的 dbSNP 中也有直接的对应内容，这些物种包括模式生物（小鼠）、宠物（猫、狗）、害虫（蚊子），以及农业上重要物种，如稻米、家畜等。dbSNP 的资料在许多格式下可以自由使用。

SNP 资料库有 4 种主要的提交类型：①SNP 的发现；②基因频率和（或）基因型频率；③个体的基因型；④单倍型。网上描述了不同类型提交的格式：http://www.ncbi.nlm.nih.gov/projects/SNP/get_html.cgi?whichHtml=how_to_submit。每个提交者在他的提交中要描述所使用的方法是用于检测变异的技术还是用于估计基因频率的技术，需要在一个独立的文本中提供技术方法的详细描述。

## SNP 的发现

在提交给 dbSNP 时，一个 SNP 的发现的基本内容应包括在多态性周围的核苷酸序列，该变异结构与基因组其他部分串在一起，以及 mRNA 序列。dbSNP 接受基因组 DNA 或 cDNA 序列（mRNA 转录本序列）的提交。提交的序列应至少有 100 bp 的长度以包括较大的参考基因组序列，以便于提高这一序列的特异性。

## 获得和建造的循环

由于 SNP 编号在讨论和稿件中被广泛用于描述基因组变异，因此我们需要描述一下编号的过程。数据库的周期性发布称为构建（build），一般来讲，dbSNP 构建的发布与每个物种各自基因组组装的连续构建过程协调一致。当基因组的构建不经常进行时，如果在两次汇编之间有一组实质性的新材料递交，dbSNP 的构建即发布。2007 年，dbSNP 进入固定的发布周期，每年 4～6 次。

每个新的构建开始于"资料的结束"，它定义为这套新的提交资料将由 MegaBLAST 组合进入基因组序列，随后进行簇的重建和注释。我们的排列启发（heuristics）的详细内容可在互联网找到：在 http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.section.ch5.ch5-s8 上的 dbSNP FAQ 上。进入每次构建的资料通常包括上次资料的结束后所有新收到的递交材料。

### 递交

一个序列变异递交到 dbSNP，其术语称之为 subSNP。递交时得到 submited sequence accesion ID，就是 ss#。这些数字通常出现在发现多态性的文章中。

### **dbSNP 的构建创建无冗余簇递交组**

递交 SNP 发现时，首先给予一个唯一的递交 SNP（subSNP）的获得号码，设计为前缀是 ss，后面跟一个整数。在一个新的构建或构建周期过程中，这些新的递交组合进入一个或多个基因组集合中，与现有的记录集合在一起，在每个靶集合中建立一套变异位置的无冗余的各变异的组合。当两个或更多的 subSNP 在作图的集合中共定位于同一套位置时就形成了一个簇。然后每个 ss 数的簇就被给予一个唯一的参考 SNP（refSNP）获得号码，设计为前缀 rs，后面跟一个整数。

### 参考 SNP 的获得

一个 dbSNP rs#，在 dbSNP 术语中称为 refSNP，在 dbSNP 构建中定义为一个唯一的作图结果。经筛选同一组排列的 SNP 递交形成一个簇，组合进一个常规的 rs# 记录。尽管我们期待具有足够的旁侧序列复杂性的递交被排列在基因组中单一的位点上，但有一个重要的例外，在低复杂性或重复的基因组序列区域的变异位置将会不明确，并

且在交替单倍型区域的变异（举例来说，HLADR1/52/53 或 X 染色体假常染色体区）将有多个正确的定位，而这种变异在基因组中仍然被记录成唯一的。

一个新的、还没有与任何已经存在的 subSNP 组合在一起的 subSNP 被给予一个新的 rs 数码。获得的号码不再被使用；然而，subSNP 可能被递交者收回，结果在下一次构建中，这些获得的号码失活。当簇里只含有单一的 subSNP 时，撤回递交也可能导致 refSNP 的失活。

应注意到获得的过程依赖于新的 ss 和已存在的 refSNP 的作图，这一点很重要。这些记录的作图结果还要依赖于潜在的装配顺序，还有作图时使用的运算法则。作为装配顺序和（或）作图算法变化的结果，refSNP 在构建循环中可被融合，只要 refSNP 被融合，新的簇就用具有较小整数的 rs 号码。使用已经作废的号码进行咨询仍然可行。

### 由所提交的侧翼序列，而不是组装过程确定“链”的聚类

提交者可以自主确定用哪一条 DNA 链定义不同变异，因此在提交 refSNP 簇时可以报告正链或负链。refSNP 的方向、序列和等位的序列由一个聚类的范本来确定。为方便起见，在提交一个聚类时采用最长的序列作为聚类的范本。在随后的构建中，一个新的提交可能含有更长的侧翼序列（因此可成为聚类范本的候选）。由于提交的侧翼序列的方向是随意的，新聚类范本的候选将有 50%的可能性与现行 refSNP 聚类范本的方向相反。这种情况发生时，我们通过使用候选聚类范本侧翼序列的反向互补来保留这个 refSNP 的定位（在 BLAST/BLAT 队列任务的 FASTA 格式中）。

一旦建立后，不管新的提交怎样，一个 rs# 聚类都有稳定的链的方向。当聚类的进程决定该聚类所有序列成员的方向时，它就将观察到的已按照链的方向规格化的 refSNP 聚类的等位基因汇合成一个组。

## 装配顺序的注释

一个 rs 聚类的重量由 dbSNP 的处理过程确定。rs 聚类的重量反映了 ss 的侧翼序列作图（高质量）的范本在基因组上位置的数量。例如，一个具有重量 1 的 refSNP 在人类的参考组装中有一个单一的高质量的排列；重量 2 是在一条染色体上有两个作图位点；重量 3 在两条染色体上有两个作图位点；重量 10 有三个或多个位点。refSNP 的重量在每次组装中被分开计算，这些组装由一组高质量的排列构成。因此，一个 rs 聚类可以在一个参考组装中是重量 1，而在另一次组装时定义成更高的重量。

我们将重量 1 和重量 2 的 refSNP 变异注释成在 NCBI RefSeq 染色体、邻近（contig）序列、mRNA 和蛋白质上被限定为复等位基因的变异特征（每个等位基因算一个）。重量 2 的特征有一个额外的警告注释以表明作图结果的模糊特征，这一结果来源于真实的相似或不完全的序列组装。我们不相信重量 3 和重量 10 的变异有足够的效能以保证其注释，但这些变异的作图结果在 dbSNP 中仍是可用的。我们将 NCBI RefSeq 染色体、contig 序列、mRNA 和蛋白质的 NoVariation 记录注释为混杂的特征（miscel-

laneous feature）或者 misc _ feat。所有的 dbSNP 的注释也包括一个返回到 dbSNP 的 db _ Xref 相互参照点，一个 refSNP ID（rs）号码和一个确认的和对该 rs 聚类链接信息位图编码的摘要。

GenBank 记录只能由最初的作者进行注释。因此当 dbSNP 的处理过程发现高质量的 refSNP 采样数录入 High-Throughput Genome Sequence（HTGS）和 GenBank 不多余的分割时，它们通过内部的 NCBI 序列注解数据库将其连接起来。这些可以通过选择在 GenBank 记录中显示的 SNP 检验盒（checkbox）而重新得到。

## 泛函分析

dbSNP 通过在 contig 注解过程中检查作为基因特性的侧翼序列来计算序列变异的功能关联。当前的构建包括在 RefSeq 和 GenBank mRNA 这一处理过程中。表 5-1 定义了 dbSNP 变异的功能分类。当一个变异在一个转录本附近，或者在一个转录本内的非编码区时，我们通过相对于转录本序列结构的变异位置确定功能分类。换句话说，一个变异可以在一个基因附近（locus region），在一个非翻译区（UTR）（mRNA-utr）内，一个内含子内，或在剪切位点上（splice-site）。如果这个变异在编码区，这个变异的功能分类依赖于每个等位基因对预期翻译的多肽序列的影响。由于功能的分类由位置和序列的参数决定，产生了两个事实：①如果因为选择性剪切使一个基因有多个转录本，那么一个变异对这个基因就可能有多种不同的功能关系。②如果许多基因紧密排列在一个连续的区域内，那么基因组中一个单一位点的变异对它的邻近基因来说就可能具有多重、潜在、不同的关系。

**表 5-1　dbSNP 变异的功能组**

| 分类号码 | 分类 ID | 分类描述 |
|---|---|---|
| 1 | locus region | 变异是在一个基因特征(在任何链)的 2kb 5′或 500bp 3′之内,但这个变异不在该基因转录本中。这一组在绘图概要中用 L 表示 |
| 2 | coding | 变异是在一个基因的编码区。这一组在绘图概要中用 C 表示 |
| 3 | coding-synon | 变异等位基因与一个基因的编码区是同义的。如果将该等位基因替换密码子并翻译时并不能造成由这一序列编码的氨基酸的变化，该等位基因就可以归于这一组。如果所有的等位基因都归为 contig 参考或 coding-synon,则这个变异是同义的替换。这一组在绘图概要中用 C 表示 |
| 4 | coding-nonsynon | 变异等位基因与一个基因的编码区是非同义的。当用该等位基因替换密码子并翻译时能造成由这一序列编码的氨基酸的变化,该等位基因就可以归于这一组。如果任何等位基因归为 coding-nonsynon,则这个变异是非同义的替换。这一组在绘图概要中用 C 或 N 表示 |
| 5 | mRNA-UTR | 变异在一个基因的转录本中,但是不在该转录本的编码区内。这一组在绘图概要中用 aT 表示 |

续表

| 分类号码 | 分类 ID | 分类描述 |
| --- | --- | --- |
| 6 | intron | 变异在一个基因的内含子中,但不是内含子的前两个或最后两个碱基。这一组在绘图概要中用 L 表示 |
| 7 | splice-site | 在内含子的前两个或最后两个碱基出现变异。这一组在绘图概要中用 T 表示 |
| 8 | contig-reference | 变异等位基因与 contig 核苷酸一致。典型的,一个变异的等位基因与参考的基因组相同。这一组在绘图概要中用 C 或 N 表示,根据这个变异替换的等位基因的情况而定 |
| 9 | coding-exception | 变异在一个基因的编码区,但由于外显子排列错误而不能确定精确位置。这一组在绘图概要中用 C 表示 |

典型的变异是双等位的——一个变异的等位基因与 contig 一致（contig 参考），另一个等位基因或是同义的或非同义的改变。在多于一个等位基因被鉴定的情况下，有可能一个等位基因是同义的改变，另一个是非同义的改变。任何非同义的改变将被归入非同义变异中；反之归入同义变异中。可能时，表达为蛋白多肽变异的 snSNP 与该蛋白质的三维结构关联。

多数基因的特征由变异相对于转录本外显子边界的位置来定义。然而对于编码区的变异，每一个等位基因有一个功能组，这是因为分组依赖于等位基因序列。

## 群体频率

二倍体生物具有两套常染色体，每套遗传自双亲之一。因此每个个体具有每个常染色体 SNP 的两个内容（等位基因）。如果检查一个个体的 G 和 T 等位基因的 SNP，可以得到三种可能的基因型：GG、GT、TT。在不同的群体中遗传变异的等位基因通常具有不同的频率。一个群体中最常见的等位基因因为种种原因有可能在另一个群体中非常稀少（甚至完全不存在）。当一个群体与其他相邻群体处于生殖隔离时，如宗教隔离群体或岛屿群体，等位基因的变异就可能成为一个独有的多态现象。

dbSNP 可用于估计 SNP 的每一个等位基因的群体频率，并可提供个体的基因型资料（见下文叙述）。频率估计可以根据进行测量的实验方法的精确度，作为等位基因计数或二进制的频率间隔而被提交到 dbSNP。dbSNP 包含了特定群体样本等位基因频率的记录，由每次提交所定义。

每次提交将一个群体（一组样品）限定为用来发现变异的一个组，或用来确定群体特异性测量等位基因频率的一个组。在一些实验设计中这些样品可以是重叠的。dbSNP 根据群体的地理起源将其归入不同样品组。这种广泛的分类方法提供了在 dbSNP 中描述样品的一般框架，但不适合于严格的群体遗传学分析。广泛的遗传信息在可能时被限定在个体水平上（见下文叙述）。与方法描述类似，群体描述需要递交者提供群体 ID 和样品组的自由体裁描述。

## 个体基因型

dbSNP 知识库用来为 HapMap 计划提供无冗余的 SNP 特性和作图信息，这一计划用由 270 个个体得来的 dbSNP 构建了 330 万个 SNP。大约有 110 万个公共基因型来自于基因组水平的基因型计划的公共-个体计划，该计划在第一阶段由 HapMap 联盟产生，第二阶段由 Perlegen 产生（图 5-1）。其他的人类基因型资料主要致力于基因的重

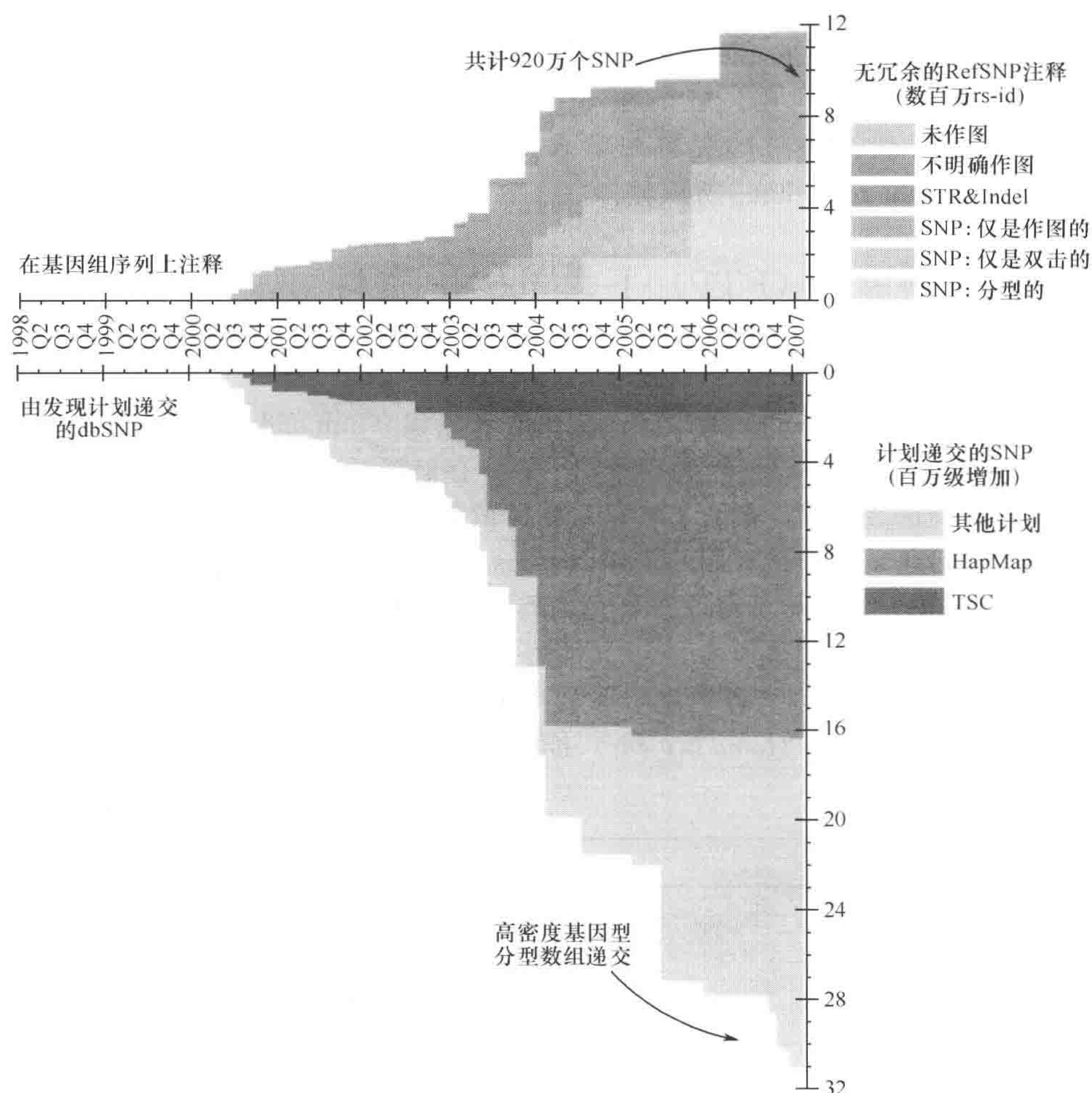



图 5-1　从 1998 年到现在人类 dbSNP 的增长。在时间标线下面的图显示自从 1998 数据库建立后提交到人类 dbSNP 的数目（用百万 ss＃'s 数表示）。两个著名的计划：1999～2002 年 SNP 联盟的 TSC 内容和 2003～2004 年的鸟枪法重新测序得来的 HapMap 内容，约占据了提交到 dbSNP 的 50％的内容。其余的人类资料由其他提交者补充。上方的图显示了无冗余的人类变异的增长，包括 230 万个插入/缺失的记录和构建 127 中的 920 万个 SNP。SNP 由浅渐深分别表示建立基因型的（genotyped）、双击的（double-hit，指普通的、但未生效的）和计算可预知的 SNP

新测序，如 PGA 和 EGP（两者在简介中都做了讨论），这些也对发现记录新的 SNP 的实际数目作出贡献。已经存放了从所有递交者得来的大约 120 万人类基因型。在可预见的未来，从人类样品中得到的基因型资料的数目将呈指数形式增长。当新的方法被评估和检测，信息被编辑以利于 SNP 标签的选择时，公共可利用的基因型作为已知样品就可以使用。

NCBI 从 2006 年 12 月开始发行基因型和表型资料（dbGaP），提供由例子/对照和纵向全基因组关联研究（whole-genome association study）这里或广泛地对照知识库递交的基因型和表型信息。这些资料与这里提到的 dbSNP 的内容不同，已被允许用于医学研究，但不能无限地扩散。因此，这些资料通过在 NIH 项目赞助下运行的一个控制途径和分配机制方可利用。可以在 http://view. ncbi. nlm. nih. gov/dbgap/上浏览 dbGaP。访问和下载被认可的资料的请求由 dbGaP 授权的访问登录系统 http://view. ncbi. nlm. nih. gov/dbgap 所控制。

## 通过文件传输协议（FTP）下载 dbSNP 资料

### FTP 位点的目录结构

自从构建 125 后，dbSNP 的设计改为中心和轮辐（hub and spoke）模型。dbSNP 的主要部分（dbSNP _ Main）作为轮子的中心，存储资源中所有的中心表；每个轮辐是生物体特异性的数据库，包含了特定生物体的最新资料。主要 dbSNP 的 FTP 目录包含一系列的目录，其中最有用的是 organisms/、database/、specs/。FTP specs/目录包括 . text、. pdf、. asn 和 . xsd 文件，上面有从 dbSNP 提交的指导、构建 125 作图信息、基因型资源信息，到单倍型 . xsd 文件的提交等所有内容。organisms/目录包括具有 dbSNP 资料的生物体列表，通过普通名称进行排列，后面跟随 NCBI 分类法的 ID 号码。例如，人类物种（*Homo sapiens*）排列在 human _ 9606。human _ 9606 的亚目录包括：

| FTP 亚目录 | 注释 |
|---|---|
| ASNl_bin/ | ASNl 二进制文件，doc sum 格式 |
| ASNl_flat/ | ASNl 文本文件，doc sum 格式 |
| XML/ | XML 文件，doc sum 格式 |
| chr_rpts/ | 从 doc sum 选择的区域，安排成 tab 界定的文件 |
| Rs_fasta/ | Fasta 格式的文件 rs 聚类，使用 SS 范本序列 |
| Submit_format/ | 递交的例子 |
| Genotype_by_gene/ | 使用 genoExchange 格式的跨基因区域的基因型 |
| genotype/ | 使用 genoExchange 格式的染色体文件 |
| haplotypes/ | 使用 hapExchange 格式提交的单倍型 |
| database/ | 数据库方案目录的指针（见下文叙述） |
| misc/ | 各种生物体特异性的文件 |

数据库/目录指示特别重要。它包含为 dbSNP 创建表格和索引所需要的方案、数据

和 SQL 陈述。

The../shared _ schema 子目录包含了针对 dbSNP _ main 而建立的方案 DDL（SQL Data Definition Language，SQL 数据定义语言）。

The../shared _ data 子目录包含了 dbSNP _ main 数据库中的数据，该数据库由所有生物物种共享。

The../organism _ schema 子目录包含了进入方案 DDL 的每个生物物种特异数据库的链接。

The../organism _ data 子目录包含了进入每一个生物物种特异的数据库相连的数据的链接。

这些数据由表格组织起来，每个表格构成一个文件。文件名按惯例为〈table-name〉.bcp.gz.。作图表格的文件名按惯例也包括 dbSNP 构建的 ID 号码和 NCBI 基因组构建的 ID 号码。例如，b127 _ SNPContigLoc _ 36 _ 2 表示在 dbSNP 构建 127 期间，这个 SNPContigLoc 表格含有作图到 NCBI contig 构建 36，版本 2 的 SNP。这些数据文件在每个表格行上有一行。每个文件中的数据范围由 tab 限定。dbSNP 使用标准的 SQL DDL 创建表格，浏览这些表格和索引。一个数据库里有许多可用于产生表格/索引、创建陈述的应用程序。

## 浏览 dbSNP 的内容

dbSNP 为人们使用网络浏览器展示和下载其内容提供了许多选项。在每个 dbSNP 网页上简单罗列所含数据和链接的做法已为更新的方法所替代。我们请读者跟随下面的例子，以深入了解 dbSNP 所包含的内容。

### 使用 SNP ID 检索

使用网络浏览器是检索 dbSNP 数据库，以获得特定 rs 号码、ss 号码、或提交者的 SNP 标识符最便利的方法。在下列例子中，我们寻找 Celera 变异 hCVl234567。

（1）打开 dbSNP 的主页 http://www.ncbi.nlm.nih.gov/SNP/，在“Search by Id on All Assemblies”下面的文本框中输入 hCVl234567。

（2）在 drop-down select box 中选择“Submitter's SNP ID”选项。

（3）点击“Search”，得到一个与请求相匹配的记录列表。

在这一例子中，只列出了一个 SNP。这一请求选项允许使用通配符 *。例如，请求“hcvl2345 *”指的是 hcvl2345 后面可跟随任何内容。使用通配符请求将得到 65 个记录。

我们通过在步骤（2）中提到的“drop-down select box”中选择合适的选项，类似的方法可用于检索 rs 或 ss 号码。像下面所描绘的，Entrez SNP 是检索 rs 号码的另一种替换方式，然而，它不能用于使用 ss 号码或提交者标识符的检索请求。

### 对基因区域 SNP 的检索

为了检索位于基因区域的 dbSNP 记录，从 Entrez gene 的主页开始检索。下面描述获得在人类 CYP2E1 基因内的变异列表的步骤，同时介绍在标准的 dbSNP 网页上可使用的选项。

(1) 打开 Entrez gene 的主页：http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene，在标有“for”的文本框内不带引号地填上“CYP2E1 [gene name] AND Human [orgn]”，然后点“go”选项。这将带你到 Entrez Gene 的 CYP2E1 基因的概览中。

(2) 点击位于该概览最右边的“Links”链接。这将显示这一记录中可使用的许多资源/概览的选项。

(3) 点击“GeneView in dbSNP”选项。这将显示一个基因-中心的 dbSNP 网页。

缺省时，编码区（cSNp）SNP 将显示出来，要显示该浏览中的所有变异，点击“In gene region”单选按钮，然后再点“refresh”选项。

GeneView 网页为使用者提供了对所有变异的详细浏览，这些变异通过 NCBI 的映像管道被绘制到基因区域。Gene Model 部分在基因组连续作图结果中列出一个或多个 refSeq mRNA，其中每一个都可以作为一个基因模型。突出的基因模型决定了基因结构和 SNP 位置，以及随后的 SNP 数据列表的绘图显示。要改变基因模型，点击位于这些资料显示上面的“View snp on GeneModd”链接。SNP 的列表按照基因组连续顺序排列，该排列可能与基因的顺序相反，并根据 SNP 功能分类不同而着不同颜色。编码区的变异包括蛋白质残基和密码子位置信息。紧跟着基因模式资料表的另一个表中列出了作图到转录位置侧翼基因组位置的变异。

## 创建一个本地的 dbSNP 拷贝

dbSNP 是一个包含上百个表格的关系数据库。dbSNP 的内容可作为一套 tab 限定的资料文件（每一个文件对应一个关系数据库列表）、实体关联图表（ERDS）和资料词典通过 FTP 下载：ftp://ftp.ncbi.nih.gov/snp/。出于安全考虑和卖主授权，dbSNP不为使用者提供直接的数据库下载。这种 dbSNP 资料检索方法最适合于大规模基因型中心。学术部门和私人部门的核心信息组可能需要将他们的资料与 dbSNP 进行整合，或者提供例行的询问选择。创建一个本地的 dbSNP 拷贝的任务可能是复杂的，应交给有经验的程序师去做。应注意保持资料与现行的 dbSNP 和（或）基因组构建的同步性。随后改变方案时需要修改密码并要有足够的空间。

### 所需要的软件

#### 有关的数据库软件

如果你正在考虑创建一个本地的 dbSNP 拷贝，你必须首先有一个特定的数据库服

务器，如 MySQL、Sybase、Microsoft SQL server 或 Oracle。NCBI 上的 dbSNP 在 MSSQL server version 2000 上运行，但使用者已经成功地在 Oracle、MySQL 以及其他软件上创建了他们自己的 dbSNP 拷贝。

#### 资料装载工具

从 dbSNP FTP 站点将资料装载到数据库中需要一个批数据装载工具，它通常伴随数据库一起安装。这类工具的一个例子是与 Sybase 在一起的 bcp（bulk copy）软件；或 MSSQL 服务器中的“bulkinsert”命令。

#### 解压软件

dbSNP 使用 winzip/gzip 来压缩/解压缩 FYP 文件。可以在 dbSNP FTP 站点找到解压缩 *.gz 和 *.Z 文件的详细指导。

### 所需要的硬件

#### 计算机平台/OS

数据库可以安装在任何有互联网连接的 PC、Mac 或 UNIX 上。

#### 磁盘空间

目前一个完整的，包含所有生物种 dbSNP 的拷贝需要 500GB 的空间。你可以根据你所感兴趣的物种创建一个只包含这些物种资料的本地数据库，应保证硬盘有继续发展的空间。

#### 内存

当前的 sql 服务器对 dbSNP 具有 4GB 的内存。

#### 互联网连接

我们推荐高速连接以便下载相当庞大的数据文件。

## dbSNP 的物理模型

用户需要在理解数据库方案的基础上创建和维持自己的 dbSNP 拷贝，这是因为数据库方案是 dbSNP 的直观表达，它显示了 dbSNP 中资料成分的逻辑关系。所有使用者都应该参考在 ..database/shared _ data/目录中的文件 db-DataDictionary.bcp.gz，和位于 ..database/目录下的 ERD 文件 erd _ dbSNR.pdf。附属文件还包括前缀“dn _ ”导出的资料表格，和前缀“db _ ”的数据库特异性资料表格，如 db _ index _ sql.bcp.gz。

#### 在线创建一个本地的 dbSNP 拷贝的过程

这些指导假定对象是有数据库相关知识的读者，而不是新手。

(1) 创建一个你所需要的数据库结构。它可包括一个或多个物种特异性的数据库，带有或不带有共享的 dbSNP _ main 数据库。由 dbSNP _ main 来的表格应该永远作为你的本地拷贝的一部分而被包括在内。

(2) 为每一个数据库创建表格，不需要约束或索引。

(3) 使用数据库工具从 . bcp 文件向表格批量插入资料。

(4) 为表格创建索引和约束。

如果你在创建本地的 dbSNP 拷贝时遇到问题，请与 snpadmin@ncbi. nlm. nih. gov 联系。

## 结构文件格式 ASNl 和 XML

资料可以通过结构文件格式 ASNl 二进位、ASNl 文本和 XML 从 dbSNP FTP 站点下载。这些文件位于 FTP 目录的物种水平的子目录下（如 human _ 9606）。对于dbSNP 构建 125，DocSum ASNl 格式直接映射在由 XML 方案限定的 DocSum XML 上，该方案在 ftp：//ftp. ncbi. nih. gov/snp/specs/docsum _ 2005. xsd 上。方案版本周期性地改变，应该使用最新的版本，可以在 ftp：//ftp. ncbi. nih. gov/snp/database/b124/mssql/schema/erd _ dbSNP. pdf 找到这一方案的图形显示。

在每次构建中一些 flat file 格式也是可用的，如染色体报告、rs 和 ss fasta、基因组报告等（这也可以在生物种的亚目录下找到）。

## 基因型的网络服务

我们通过描述如何使用用于查询基因型信息的网络服务来归纳关于 dbSNP 的讨论。对于构建 126，可以直接使用网络浏览器和基因型服务器的网页，或者程序化地直接与服务部门通话来询问 dbSNP 基因型的信息。

### 使用网络浏览器进行基因型的询问

使用 http://www. mcbi. nlm. nih. gov/projects/SNP/snp _ gf. cgi 上的基因型询问方式，使用者通过一个简单的、互动的方式访问和下载一小套基因型（少于 20 000 基因型）和等位基因频率报告。单独的基因型和等位基因频率的信息可通过三个步骤得到，当前的步骤被粗略显示在网页的上方，跟随着一个或多个动作按钮。第一个步骤允许使用者选择物种和基因型的显示方位，其方位可以是参考基因组的方位或是 rs 的方位。然后使用者通过点击“Region”、“RS Numbers”或“Gene”选项分别得到将变异限定在基因组区域的结果，一个 rs 数目的列表或一个基因区域的选项。报告可使用三种输出格式：html、tab-delimited text 和 xml。以下的事例说明了许多基因型服务器选项的使用方法。

事例 1：用 rs 的方位创建一个基因型报告，说明小鼠 *IL6* 基因上的两个特异 SNP。

(1) 在物种处选择小鼠（10090）、显示方位和 RS 缺省，单选按钮。

（2）点击 RS 号码 tab 并在文本区敲入 rs8259847 rs8259848。

（3）点击下一个钮进入步骤（2）选择群体。

（4）通过点击框中的加号扩大群体树并寻找“ROCHBIO”复选框。

（5）点击“Download XML”进行 xml 输出。你可能不得不让浏览器明确地显示或下载文件，视你的浏览器设置而定。

事例 2：用基因组方位创建基因型报告，说明排除了内含子 SNP 的人 *ABO* 基因，限定在欧洲裔群体。

（1）使用“Species human（9606）”、“Display orientation”和“Genome-default”的默认选择，单选按钮。

（2）在“Enter the Entrez Symbol or ID of a gene”文本框点击“Gene tab”并输入 ABO。

（3）检查“Weight 1”以找到“Limit Search To...”。

（4）从“SNP Function Class”选择框选择“intron”并点击位于选择下方的“invert”。点击“invert”可反向选择在选择框中的选项，点击“invert”再点击“all”将清除所有的选择。

（5）点击“Next”钮进入步骤（2）“Select Populations”。

（6）核对“European”。默认时各群体将根据宽泛的遗传组合起来。事例 3 表明怎样通过提交者 Handle 组合各群体。

（7）点击“Display Results button”以输出 html。

事例 3：用 rs 方位创建一个基因型报告，说明人类第 6 号染色体，位于 31 350 000 和 31 430 000，将结果限定在 HapMap YRI 群体内。

（1）使用“Species human（9606）”、“Display orientation”和“Genome-default”默认选择，单选按钮。

（2）点击“Region tab”并在“Chromosome”、“From”和“ To”文本框分别输入 6、31 350 000 和 31 430 000。

（3）对“SNP Properties section Weight 1 checked”和没有“SNP Function Class”的选择使用默认选择。

（4）点击“Next”钮进入步骤（2）“Select Populations”。

（5）从“View population by”选择框中选择“Handle...”。

（6）点击“CSHL-HAPMAP”旁边的“boxed plus sign”以扩大可利用的群体。

（7）点击“on HapMap-YRI”检验框。

（8）点击“Download Text”，以文本输出单个的基因型。你可能不得不让浏览器明确地显示或下载文件，视你的浏览器设置而定。

使用能够访问和翻译 http 请求的程序或原本也可以访问基因型服务器。表 5-2 列举了可以用来修正服务器请求的 http 有用参数。样品原本 1 说明了 dbSNP 使用在 Perl 上 的 NCBI EUtils 功能对 refSNP 内容的询问。样品原本 2 提供了一个可以被用来取得基因型和等位基因频率报告的例子。表 5-3 列举了服务器支持的返回报告的类型。

表 5-2　dbSNP 基因型网络服务器识别的 HTTP 参数

| 名字 | 页码 | 值 |
| --- | --- | --- |
| pg | 所有 | 第一页、第二页、第三页分别为 0、1、2 |
| species | 第一页 | 搜索物种或 tax ID。可以是一个整数形式的 tax ID 或先写一个物种名，跟着下划线和 tax ID。例如，human_9606 和 9606 是等同的 |
| tax id | 第一页 | 与物种相同，如果物种已经特指可忽略掉 |
| RSPick | 第一页 | 通过 RS 数码选择 SNP 时选 1，通过名称或 ID 选择基因时选 2，其他的为空 |
| pickRS | 第一页 | 由空格、逗号、换行限定的 rs 号码列表。只有当 RSPick 是 1 时使用 |
| gene | 第一页 | 基因的名称或 NCBI ID，只有当 RSPick 是 2 时使用 |
| chr | 第一页 | 染色体。可以是一个数目或一个字串，如 X、Y、Un 或 MT。只有当 RSPick 不是 1 的时候才使用它 |
| rng | 第一页 | 待选择的 SNP 的范围，有效值是 C 和 F。F 规定范围，C 规定该范围在 RS 数码或碱基对中居中。只有当 RSPick 不是 1 的时候才使用它。更详细的内容见下文叙述 |
| rngspec | 第一页 | 值是 chr_pos 或 rs。如果是 chr_pos，那么范围将通过在染色体内的碱基数选择；另外，根据 rng 的值，其范围也可以在两个 RS 数目之间或在一个特定的 RS 的 bp 数目之内。只有当 RSPick 不是 1 的时候才使用它 |
| from | 第一页 | 这是用 bp 表示的染色体上的位置，或是一个提示开始或范围的中心的 RS 数，依 rng 的值(见下文叙述)而定。RS 数可选择地前置一个 rs，bp 量可选择地用 k(代表千)或 M(代表百万)结尾。只有当 RSPick 不是 1 的时候才使用它 |
| to | 第一页 | 如果 rng 是 F，这就是一个结束 RS 码或 bp，使用与 from 参数同样的格式。如果 rng 是 C，这就是一个由 from 参数确定的从中央位置开始的 bp 数。只有当 RSPick 不是 1 的时候才使用它 |
| weight 1 | 第一页 | 仅在选择重量 1SNP 时选 1，其他情况下省略。只有当 RSPick 不是 1 的时候才使用它 |
| trusnp | 第一页 | 仅在选择 SNP 时选 1，其他情况下省略。只有当 RSPick 不是 1 的时候才使用它 |
| diall | 第一页 | 仅在选择两对等位基因的 SNP 时选 1，其他情况下省略。只有当 RSPick 不是 1 的时候才使用它 |
| snpfnc | 第一页 | SNP 函数。一个代表 SNP 函数的整数。可以选择多值，但它们应是不同的名字=数值，如 snpfnc=2、snpfnc=3、snpfnc=4 等。该函数可选择但只有当 RSPick 不是 1 的时候才使用它。表 5-1 给出了值 |
| founders | 第二页 | 如果选择的建立者在系谱内选 1，其他情况下省略 |
| pop | 第二页 | 对于人类(9606)：群体和类型。在群体表上 pop_id 的整数字串，随后跟随一个解释要点(!)和整数的群体组。这些群体组包括：<br>0——未定义的<br>1——非裔美国人<br>2——亚洲人<br>3——欧洲人<br>4——全球的<br>5——西班牙裔的<br>6——本土美国人<br>7——北非/中东<br>8——撒哈拉以南非洲人<br>9——交织的情况<br><br>对于非人类物种，只有一个整数 pop_id。<br><br>可以使用多重名字=数值对。 |
| ind | 第二页 | 只用于非人类。在 SubInd 表的 Submitted_ind_id。可以使用多重名字=数值对 |

续表

| 名字 | 页码 | 值 |
| --- | --- | --- |
| type | 第二页,第三页 | 下载 xml 时的值是 xml,标记分开的文本时是 text,恢复 html 时是 empty 或 omitted |
| reportId | 第三页 | 一个报告 ID,用于进一步询问,直到资料备齐。如果这些资料存在,使用 reportId 和 pg=2 的询问可以取得它们,否则它将再次显示包含 reportId 的网页。当使用浏览器时,这一请求每 5 s 重复一次,直到找到资料。 |
| api | 第三页 | 当寻回的资料处于自动状态时,在原本总是包括这些伴有值 1 的参数。这一反应或者是资料,或者是有一个标签的 XML,带有时域号码报告 Id,用于随后的请求和单一的特征状态。这些状态的值如下：<br><br>0——没有资料<br>1——已有资料<br>-1——reportId 无效或过期<br><br>随后的请求应该带有如下的参数：<br><br>reported=session nimber retrieved<br>pg=2<br>api=1<br>type=xml 或 type=text<br>status=1（可选择的,见下文叙述）<br><br>发送一个时域号码的目的是防止当发送大量请求时服务器过时 |
| status | 第三页 | 如果你只想得到一个报告 ID 和状态,当寻回的资料处于自动的状态时,在原本中总是包括这些伴有值 1 的参数。这样做是为了防止分列结果,以判断你是在接收这个报告还是仅仅处于某种状态 |

**表 5-3 用 NCBI Efetch 询问 EntrezSNP 的 Efetch 报告参数**

| Rettype | 报告类型描述 |
| --- | --- |
| Brief | Docsum Brief（Entrez 默认） |
| FIT | Flat File |
| ASNl | ASN. 1 |
| XML | XML |
| FASTA | FASTA |
| RSR | RS 聚类报告 |
| ssexemplar | SS 范本列表 |
| CHR | 染色体报告 |
| GENB | 基因型 |
| GEN | 基因型细节 |
| GENXML | 基因型 XML |
| DocSet | 摘要 |
| FREQXML | 频率 XML |
| uilist | UI(RS)列表 |
| MergeStatus | RS 合并状况 |

可利用报告返回类型密码显示 EntrezSNP 询问结果。作为在 PERL 编码的例子，请见样品原本 1（见下文）

**Sample Script 1: Using Efetch to Query dbSNP and Return Results**

```
#!/usr/local/bin/perl -w
# ===========================================================================
#
#                              PUBLIC DOMAIN NOTICE
#                National Center for Biotechnology Information
#
#  This software/database is a "United States Government Work" under the
#  terms of the United States Copyright Act.  It was written as part of
#  the author's official duties as a United States Government employee and
#  thus cannot be copyrighted.  This software/database is freely available
#  to the public for use. The National Library of Medicine and the U.S.
#  Government have not placed any restriction on its use or reproduction.
#
#  Although all reasonable efforts have been taken to ensure the accuracy
#  and reliability of the software and data, the NLM and the U.S.
#  Government do not and cannot warrant the performance or results that
#  may be obtained by using this software or data. The NLM and the U.S.
#  Government disclaim all warranties, express or implied, including
#  warranties of performance, merchantability or fitness for any particular
#  purpose.
#
#  Please cite the author in any work or product based on this material.
#
# ===========================================================================
#
# Author:  Michael L. Feolo and Lon Phan
#           Staff Scientists
#           National Center For Biotechnology Information
#           National Library of Medicine
#           National Institutes of Health

#-------------------------------------------------------------------------
# EUTIL ACCESS TO dbSNP FOR PROGRAMMATIC DATA RETRIEVAL
#-------------------------------------------------------------------------
use LWP::Simple;
my $utils = "http://www.ncbi.nlm.nih.gov/entrez/eutils";
my $db      = 'snp';

#-- Three sample queries. Replace my $term to explore.
# Sample query term to retrieve nsSNPs in a human gene (OXTR)
# my $term   = 'OXTR AND human';

# sample query term to retrieve SNPs in a 2MB interval on chromosome 1
# restricted to SNPs with genotype results (any organism)
#my $term   = '8810000:8830000[CHRPOS] AND 1[CHR] AND TRUE[GTYPE] AND Human[ORGAN-
ISM]';

#-- Sample query to retrieve SNPs in human ABO gene with genotypes.

my $term = 'TRUE[GTYPE] AND Human[ORGANISM] AND ABO[GENE]';

# report options uilist, XML, ASN.1, FASTA, etc.
# addition reports and eUtils help are online:
# http://www.ncbi.nlm.nih.gov/projects/SNP/SNPeutils.htm
# Refer to Table 3 to see current list of report type options.

my $report = 'XML';

#-- replace $report with the following values to explore the various types of
reports.

#my $report = 'FASTA';
#my $report = 'uilist';
#my $report = 'GENXML';

# ------------------------------------------------------------------------
# $esearch contains the PATH & parameters for the ESearch call
# $esearch_result contains the result of the ESearch call
```

```
# the results are displayed and parsed into variables
# $Count, $QueryKey, and $WebEnv for later use and then displayed.

my $esearch = "$utils/esearch.fcgi?" .
    "db=$db&retmax=1&usehistory=y&term=$term";

my $esearch_result = get($esearch);

print "\nESEARCH RESULT: $esearch_result\n";

$esearch_result =~
m|<Count>(\d+)</Count>.*<QueryKey>(\d+)</QueryKey>.*<WebEnv>(\S+)</WebEnv>|s;

my $Count     = $1;
my $QueryKey  = $2;
my $WebEnv    = $3;

print "Count = $Count; QueryKey = $QueryKey; WebEnv = $WebEnv\n";

# -------------------------------------------------------------------------
# this area defines a loop which will display $retmax citation results from
# Efetch each time the Enter Key is pressed, after a prompt.

my $retstart;
my $retmax=1000;


for($retstart = 0; $retstart < $Count; $retstart += $retmax) {
  my $efetch = "$utils/efetch.fcgi?"
               "rettype=$report&retmode=text&retstart=$retstart&retmax=$retmax&"
               "db=$db&query_key=$QueryKey&WebEnv=$WebEnv";

  print "\nQUERY=$efetch\n\n";

  my $efetch_result = get($efetch);

  print "$efetch_result\n\n-----PRESS ENTER!!!-------\n";
  <>;
}
```

## Sample Script 2: Query dbSNP Genotype Server Using HTTP Requests

```
#!/usr/bin/perl
# ===========================================================================
#
#                          PUBLIC DOMAIN NOTICE
#               National Center for Biotechnology Information
#
#  This software/database is a "United States Government Work" under the
#  terms of the United States Copyright Act.  It was written as part of
#  the author's official duties as a United States Government employee and
#  thus cannot be copyrighted.  This software/database is freely available
#  to the public for use. The National Library of Medicine and the U.S.
#  Government have not placed any restriction on its use or reproduction.
#
#  Although all reasonable efforts have been taken to ensure the accuracy
#  and reliability of the software and data, the NLM and the U.S.
#  Government do not and cannot warrant the performance or results that
#  may be obtained by using this software or data. The NLM and the U.S.
#  Government disclaim all warranties, express or implied, including
#  warranties of performance, merchantability or fitness for any particular
#  purpose.
#
#  Please cite the author in any work or product based on this material.
#
# ===========================================================================
#
# Author:  Douglas J. Hoffman
#          Contractor
#          National Center For Biotechnology Information
```

```
#           National Library of Medicine
#           National Institutes of Health
#     \
# File Description:
#
#    Example Perl script to retrieve a genotype report in XML
#

use strict 'vars';
use IO::Socket::INET;
use XML::SAX::PurePerl;

package XmlHandler;

sub new
{
  my $class = shift;
  my $self = {};
  bless ($self,$class);
  $self;
}


sub start_document
{
}

sub end_document
{
}

sub start_element
{
  my ($self,$properties) = @_;
     my $name = $properties->{'Name'};
  if($name eq "reportId")
  {
    $self->{inReportId} = 1;
    my $attr = $properties->{'Attributes'};
    my $status = $attr->{'{}status'};
    my $nStatus = $status->{'Value'};
    $self->{status} = $nStatus ? $nStatus : 0;
  }
  elsif ($name =~ m/error/i)
  {
    die("Error in server request");
  }

}

sub end_element
{
  my ($self,$properties) = @_;
     my $name = $properties->{'Name'};
  if($name eq "reportId")
  {
    $self->{inReportId} = undef;
  }
}
sub characters
{
  my ($self,$properties) = @_;
  if($self->{inReportId})
  {
    $self->{reportId} .= $properties->{'Data'};
  }
}
sub entity_reference
{
  my ($self,$properties) = @_;
  my $sChar = '&' . $properties->{'Name'} . ';';
  my $data = {'Data' => $sChar};
```

```
    $self->characters($data);
  }

  sub comment
  {
  }

  sub processing_instruction
  {
  }

  1;

  package main;

  my $sHost = "www.ncbi.nlm.nih.gov";
  my $sPort = 80;
  my $sScript = "/projects/SNP/snp_gf.cgi";

  sub esc
  {
    sub hexFunc
    {
      my $c = shift;
      my $rtn = "%" . sprintf("%02x",ord($c));
      $rtn;
    }

    my $str = shift;
    $str =~ s|[^ \w\./]|&hexFunc($&)|ge;
    $str =~ s/ /+/g;
    $str;
  }

  sub BuildContents
  {
    my $hRequest = shift;
    my $contents = "";
    my $s;
    my $sName;
    my $sValue;
    my $bNotEmpty = undef;
    my @aValues;
    for $s (keys %$hRequest)
    {
      $sName = &esc($s);
      $sValue = $hRequest->{$s};
      if(length($sValue))
      {
        @aValues = split /\t/,$sValue;
      }
      else
      {
        @aValues = ("");
      }
      for $sValue (@aValues)
      {
        $bNotEmpty && ($contents .= '&');
        $bNotEmpty = 1;
        $contents .= $sName;
        $contents .= "=";
        $contents .= &esc($sValue);
      }
    }
    $contents;
  }

  sub RunRequest
  {
    my $hRequest = shift;
    my $contents = &BuildContents($hRequest);
    my $nLen = length($contents);
```

```
  my $sHTTP =
    "POST ${sScript} HTTP/1.0\r\n" .
    "Host: ${sHost}:${sPort}\r\n" .
    "User-Agent: Mozilla/5.0 (Perl Script)\r\n" .
    "Connection: close\r\n" .
    "Content-Type: application/x-www-form-urlencoded\r\n" .
    "Content-Length: ${nLen}\r\n" .
    "\r\n${contents}\r\n";
  my $io = IO::Socket::INET->new
    (PeerHost => $sHost, PeerPort => $sPort, Proto =>"TCP");
  $io || (die "Could not connect to ${sHost}\n");
  $io->print($sHTTP);
  my $bDone = 0;
  my $line = 1;
  my $lines = [];
  while($line)
  {
    $line = $io->getline;
    $line ||
      (die "Did not receive entire header from ${sHost}\n");
    $line =~ s/[\r\n]//g;
    push @$lines,$line;
  }
  ($#$lines > -1) ||
    (die "Did not receive any header from ${sHost}\n");

  $line = $lines->[0];
  ($line =~ m/ 200 OK/) ||
    (die
     "Error in response header - request was as follows:\n\n${sHTTP}\n\n");
  [$io,$lines];
}

sub CopyHash
{
  my $h = shift;
  my $hRtn = {};
  my $k;
  for $k (keys %$h)
  {
    $hRtn->{$k} = $h->{$k};
  }
  $hRtn;
}

sub Run
{
  my $hRequestIn = shift;
  my $nHeaders = shift;
  my $done = 0;
  my $line;
  my $line2;
  my $str;
  my $rq;
  my $hRequest = &CopyHash($hRequestIn);
  my $sType = $hRequest->{type};
  $hRequest->{status} = 1;
  $hRequest->{api} = 1;
  $hRequest->{pg} = 2;
  while (!$done)
  {
    $rq = &RunRequest($hRequest);
    my $xml = new XmlHandler;
    my $sax = new XML::SAX::PurePerl->new(Handler => $xml);
    my $headerLines = $rq->[1];
    my $io = $rq->[0];
    my $contents = "";

    while($line = $io->getline)
    {
      $contents .= $line;
    }
```

```
        $io->close;
        $sax->parse(Source => { String => $contents });
        my $status = $xml->{status};
        my $reportId = $xml->{reportId};
        $hRequest =
        {
         reportId => $reportId,
         api => 1,
         keepAlive => 200,
         type => $sType,
         pg => 2
        };

        if($status > 0)
        {
          $done = 1;
          $rq = &RunRequest($hRequest);
          $io = $rq->[0];
          $headerLines = $rq->[1];
          if($nHeaders)
          {
            local $, = "\r\n";
            local $\ = "\r\n";
            print @$headerLines;
          }
          while($line = $io->getline)
          {
            print $line;
          }
        }
        elsif(!$status)
        {
          $hRequest->{status} = 1; ## explicitly get status only on next loop
          sleep 2;
        }
        else
        {
          print STDERR "Genotype report is not available\n";
          $done = 1;
        }
    }
    0;
}

#
#     three example structures for "sub Run" above
#

my $hExampleRange =
{
 type => "xml",
 species => "human_9606",
 RSPick => "",
 chr => 6,
 rng => "F",
 rng_spec => "chr_pos",
 from => "30000000",
 to =>    "30050030",
 weight1 => 1,
 pop => "904!3\t902!3\t1409!3\t1371!3", ## tab separated list

};

my $hExampleRS =
{
 type => "xml",
 species => "human_9606",
 RSPick => "1",
 RSlist => "8176742 8176740 8176739 8176721 8176720 512770",
```

```
 pop => "904!3\t902!3\t1409!3\t1371!3" ## tab separated list
};

my $hExampleGene =
{
 type => "xml",
 species => "human_9606",
 RSPick => "2",
 Gene => "brca1",
 weight1 => 1,
 pop => "904!3\t902!3\t1409!3\t1371!3" ## tab separated list
};

##
## use one of the 3 examples above
##

&Run($hExampleRange,0);
0;
```

# 致谢

本工作得到国家医学图书馆和国家健康研究所内部研究项目的部分资助。

# 参考文献

Cardon L.R. and Bell J.I. 2001. Association study designs for complex diseases. *Nat. Rev. Genet.* **2:** 91–99.

Carlson C.S., Eberle M.A., Kruglyak L., and Nickerson D.A. 2004. Mapping complex disease loci in whole-genome association studies. *Nature* **429:** 446–452.

Hirschhorn J.N. and Daly M.J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6:** 95–108.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320. [http://www.hapmap.org]

Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., and Sirotkin K. 2001. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29:** 308–311.

# 互联网资源

**http://pga.gs.washington.edu/** The SeattleSNPs PGA is focused on identifying, genotyping, and modeling the associations between single-nucleotide polymorphisms (SNPs) in candidate genes and pathways that underlie inflammatory responses in humans. SeattleSNPs is funded as part of the National Heart, Lung, and Blood Institute's (NHLBI) Programs for Genomic Applications (PGA).

**http://www.ncbi.nlm.nih.gov/projects/SNP/get_html.cgi?whichHtml=how_to_submit** NCBI database of single nucleotide polymorphisms (SNPs). Guidelines for dbSNP submission process.

**http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.section.ch5.ch5-s8** The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation.

# 6 使用 HapMap 网站

Albert Vernon Smith

*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724; Genthor ehf., 101 Reykjavik, Iceland, and Icelandic Heart Association, 201 Kopavogur, Iceland*



## 简介

国际人类基因组单体型图计划（International HapMap Consortium 2005）的首要目的是产生一个人类基因组的单体型图以描述遗传变异的常见类型，加速对人类疾病的遗传起因的研究。在该计划中，来自世界范围内 4 个群体中 270 个个体的大约 390 万明确的 SNP 已作基因型分型。该计划的数据在 HapMap 网站 http://www.hapmap.org（Thorisson et al. 2005），公众可自由使用。这一站点最初是该计划产生的基因型数据的入口，现提供数据的大量下载、交互式资料浏览和别的地方无法提供的分析工具。

本章描述了该站点以及用于浏览、恢复和分析计划数据的工具。提供了如何进行几种有用的和流行的任务的细节。内容包括对获取基因型和频率数据的指导，取得 tag-SNP 用于遗传关联分析，浏览图形化的各种单倍型，下载分项的基因型数据和检验标记-标记连锁不平衡（LD）类型等。

## 方案一　使用基因组浏览器浏览 HapMap 数据

对人类疾病的遗传因素的研究通常集中在连锁和（或）关联研究，以及从被怀疑与特殊疾病进程相关的通路研究中鉴定候选基因。在研究候选基因时，研究者希望知道是否有任何常见 SNP 位于附近，这些 SNP 的等位是什么，这些等位在群体中的相对频率

有多少。研究者也对编码 SNP 特别感兴趣，其等位基因改变了基因产物的氨基酸序列，因此可能代表了功能的变异。

# 方法

## 发现和浏览感兴趣的区域

HapMap 网站的基因组浏览器提供了基因组从小型到中型区域的入口，以进行这一类型的互动开发。这一基本方案显示了如何开始使用基因组浏览器。

（1）使用任何现代网络浏览器，打开 www. hapmap. org。

（2）点击位于 hapmap. org 主页“Project Data”区下面的“Browse Project Data”链接。这将带你到 GBrowse 包的基因组浏览器（图 6-1）。

（3）查找“Landmark or Region”搜索框，输入搜索项。可以是以下任何类型的搜索项。

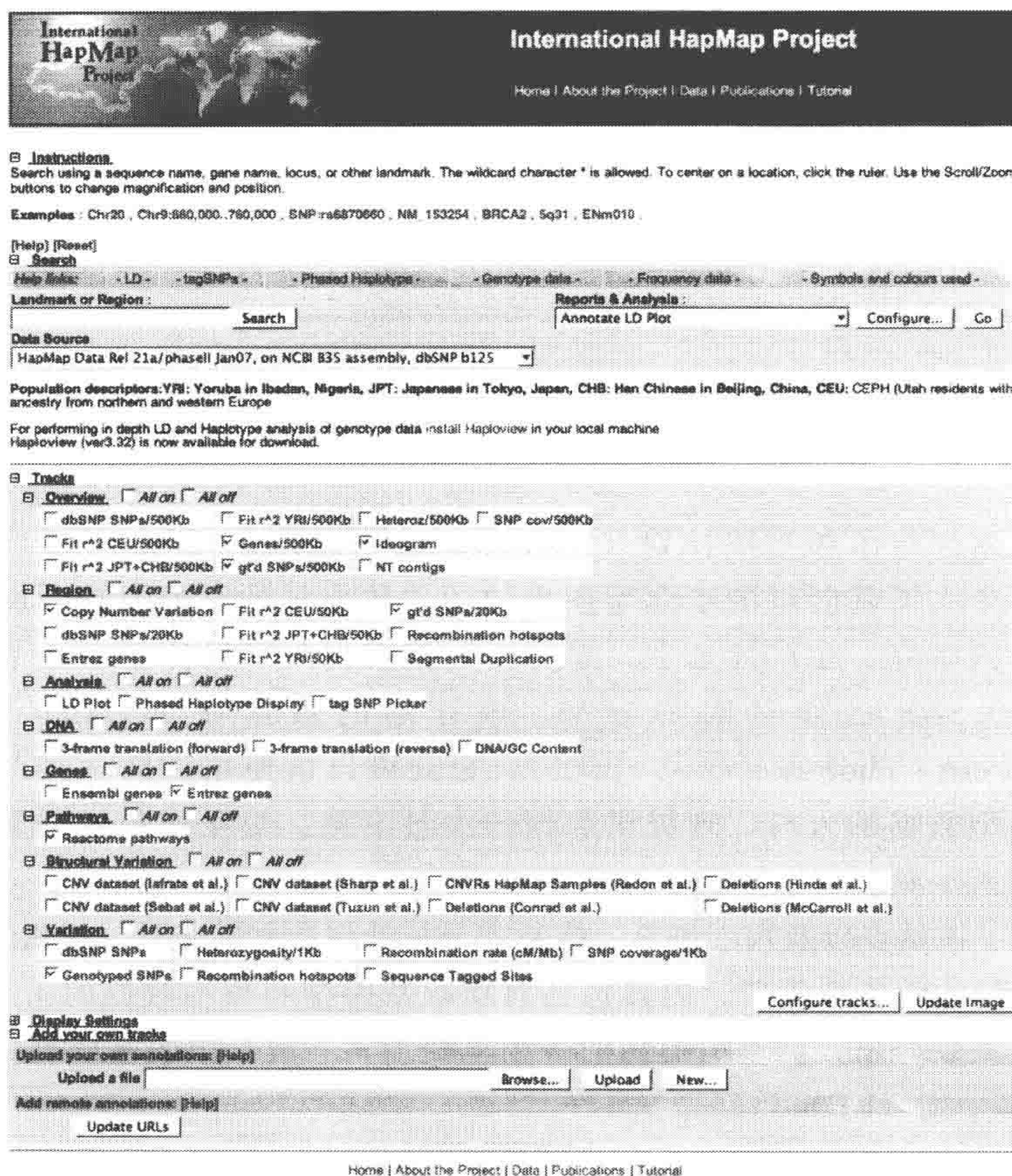



图 6-1　使用 HapMap 基因组浏览器时起始页所显示的内容。图 6-1 中是使用 HapMap 基因组浏览器时起始页所显示的内容。根据你的计算机的语言设置，本页可显示几种语言中的一种，这里显示的是英语。本页也可点击 http://www. hapmap. org/cgi-perl/gbrowse 直接到达

- 一个染色体名（如 Chr19）。
- 一个染色体部分，使用格式：染色体：开始..结束（如 Chr10：25 000..300 000）。
- SNP 的名字，使用 dbSNP rs 名（如 rs6 870 660）。
- 一个基因，使用它的 NCBI RefSeq 登录号（如 NM 153 254）。
- 一个基因，用它的普通名字（如 BRCA2）。
- 一条染色体带（如 5q31）。

(4) 进入上述界标之一后，点击“Search”钮（或者按“Enter”键）。这将回到显示围绕着所要求的特征的内容的网页（图 6-2）。如果多重特征匹配，网页将显示一个包括所有可能特征的图形摘要（包括基因组位置），提示你选择一个。

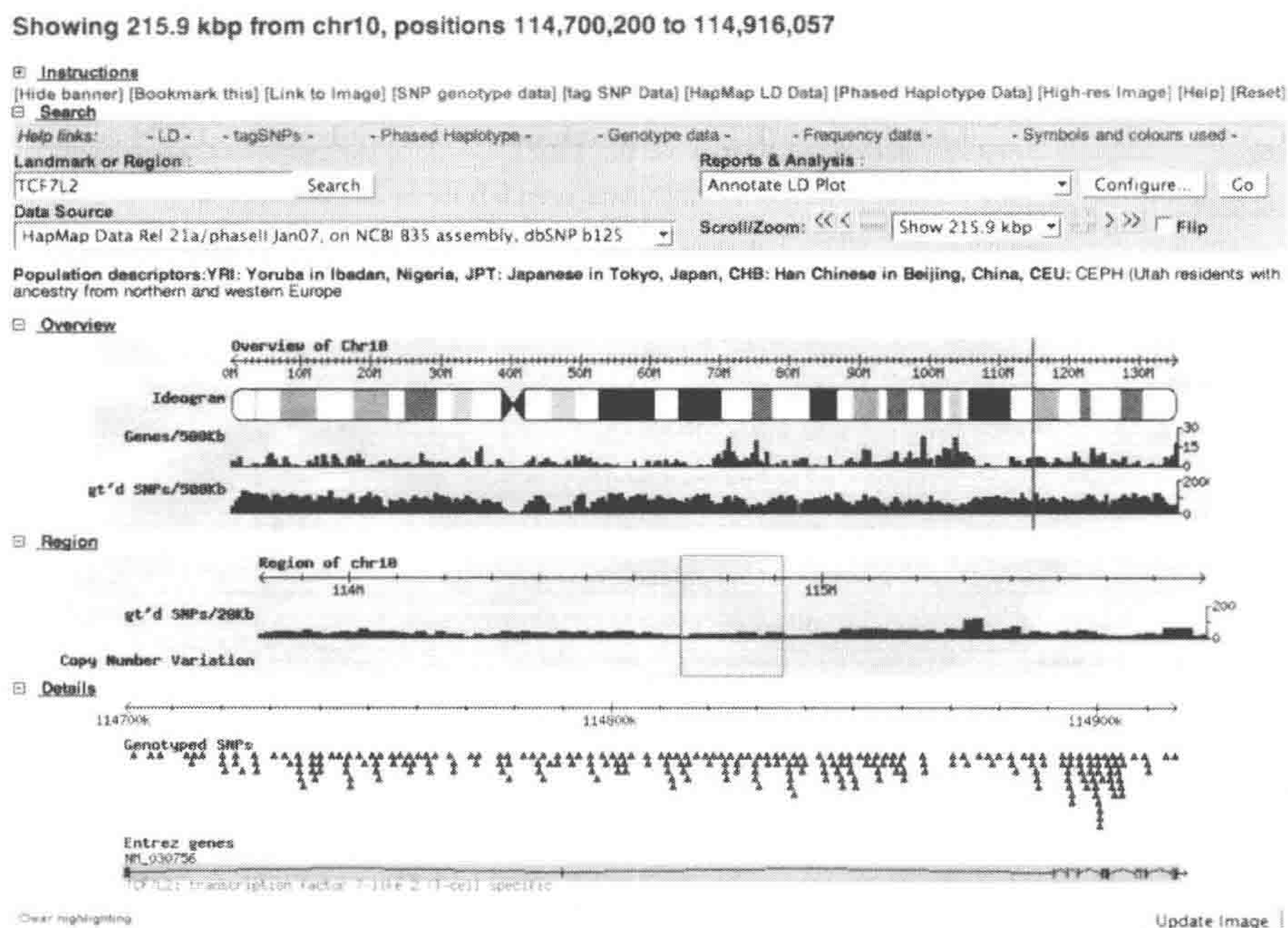



图 6-2 HapMap 基因组浏览器显示一个被请求的特征

缺省情况下，基因组浏览器显示最近发布的 HapMap 资料。通过这一界面可获得以前的资料，可以在“Data Source”菜单中选择以前发布的不同资料。

i. 在返回的网页的上面是一个“Overview”（一般观察）部分，显示的是被选择的染色体的细胞遗传学图。方框指示被考虑的染色体部分。

ii. 在此之下是一个 Region 一般观察，显示感兴趣区域周围的 2Mb。方框也指示被考虑的染色体部分。

iii. 在此之下是一个“Detail”（详细）部分，有水平的条目（track）显示不同类型的资料。缺省情况下，最初该区域只显示少量的基因组条目。两个最有用的条目，一个是“Genotyped SNPs”（基因型 SNP）条目，提供位置信息、等位基因和每个由 HapMap 计划确定的 SNP 的等位基因频率；另一个是“Entrez genes”条目，显示人类编

码蛋白质基因的位置和结构。

一些额外的信息条可供使用，这对关联研究的理解和设计可能特别有用。可利用它们对 HapMap 资料以及外部的资料来源进行一系列分析（表 6-1）。特别值得注意的是一系列与基因组的结构变异有关的条目和链接到 Reactome 的数据库（http://www.reactome.org 的条目；Vastrik et al. 2007），这是人类生物学核心通路和反应的另一个资源。

**表 6-1 在基因组浏览器中可利用的条目**(2007 年 2 月的内容)

| 种类 | 条目 |
|---|---|
| HapMap | LD plot |
| 工具 | phased haplotype display |
| | tag-SNP picker |
| 基因 | ensembl genes(Hubbard et al. 2007) |
| | entrez genes(Wheeler et al. 2007) |
| 通路 | reactome pathways(Vastrik et al. 2007) |
| 结构变异 | copy number variation (CNV) data sets(Iafrate et al. 2004;Sebat et al. 2004 |
| | sharp et al. 2005;Tuzun et al. 2005) |
| | deletions (Conrad et al. 2006;Hinds et al. 2006;McCarroll et al. 2006) |
| 变异 | dbSNP SNPs (Wheeler et al. 2007) |
| | heterozygosity/1kb |
| | recombination rate (cM/Mb) |
| | SNP coverage/1kb genotyped SNPs |
| | recombination hot spots |
| | sequence tagged sites(Wheeler et al. 2007) |
| | fit $r^2$ in genomic intervals (Smith et al. 2005) |

（5）使用网页上的控制向左或向右滚动，或改变区域的放大倍数。你也可以点击“Overview”、“Region”或在“Details”部分上面的刻度的任何部分使这一位置的内容居中。基因型 SNP 条将改变其外貌以适合图像的尺度。

i. 在低放大倍数时，基因型 SNP 被显示成等边三角形。这些颜色可通过选择“Reports and Analysis”菜单中的“Highlight SNP Properties”（加亮 SNP Properties）项加以调整。

ii. 在高放大倍数时，基因型 SNP 变为显示与 SNP 相关的等位基因。被显示为蓝色的等位基因是此位置上的参考基因组序列的等位基因，红色等位基因是 SNP 另一个等位基因。

iii. 继续放大后，基因型 SNP 条目变为显示圆形分格统计图表，显示每个基因型群体的基因频率。圆形分格统计图表的蓝色楔入显示出现在参考基因组序列的等位基因频率，红色楔入显示另一个等位基因频率。

圆形分格统计图表使研究者能方便地区分所有 4 个 HapMap 群体中高度多态的 SNP，并且这些 SNP 很有可能在其他群体中也是多态的。反之，研究者可以鉴定在单一群体更具多态性的 SNP，并且这类 SNP 适合于在群体特异性遗传筛查时作为标记使用。

（6）点击为个体 SNP 而浮起的字，查看一个文本页，上有详细的基因型、等位基因数和分析信息。

这为研究者提供产生一个 SNP 分析所必需的信息，包括创建 PCR 引物所需的左边或右边的侧翼序列。

i. 点击超文本链接到 dbSNP（http://www.ncbi.nlm.nih.gov/SNP ；Wheeler et al. 2007）以取得关于此 SNP 是怎样首次被发现的和任何其他群体遗传的更多信息，这些信息可能存在于 HapMap 项目之外。

ii. 点击到 Ensembl（http://www.ensembl.org ；Hubbard et al. 2007）的链接以到达一个网页，这里可以查到 SNP 对编码序列、拼接位点和附近基因的其他特点的结构影响。

## 检查连锁不平衡的程度

当一个研究者设计了一个课题以检测一个基因的一般等位基因变异和感兴趣的疾病的相关性时，关于这一区域 LD 程度的知识对于减少该区域需做基因型分型的 SNP 的数目而言是必需的。如果这一区域有高度的 LD，那么只需要少数的 SNP 进行基因型分型即可，这是因为它们与这一区域其他 SNP 的连锁将作为非特异性的 SNP 基因型的替代。反之，在一个低 LD 的区域，因为基因型 SNP 的等位基因状态很难预测没作基因型分型的 SNP 状态，需要更多检测。在 HapMap 计划研究的群体中确定 LD 构成是这一计划的主要目标之一。HapMap 计划在基因型 SNP 之间有预先计算好的 LD 构成。这些数据可以使用 HapMap 基因组浏览器从 HapMap 网站或从交互的浏览位置成批下载。后一种方法可以使研究者看到与感兴趣的基因分布有关联的 LD 类型。

（7）为观察从 HapMap 基因型预先计算好的 LD 资料，浏览感兴趣的区域［步骤（4）］。

（8）从“Reports and Analysis”菜单选择“Annotate LD plot”进入。

（9）点击“Configure”选项，调出配置页，这将允许你调整显示特性以适合于你。

这一页关键的参数是要显示的 HapMap 群体、使用哪种 LD 方法（D′、$r^2$ 或 LOD 的选择）、三角图是否应有方向、其顶点指向上或下、色彩设计、图中框的大小是否应该与标记之间的基因组距离成比例或是相同大小等（图 6-3）。

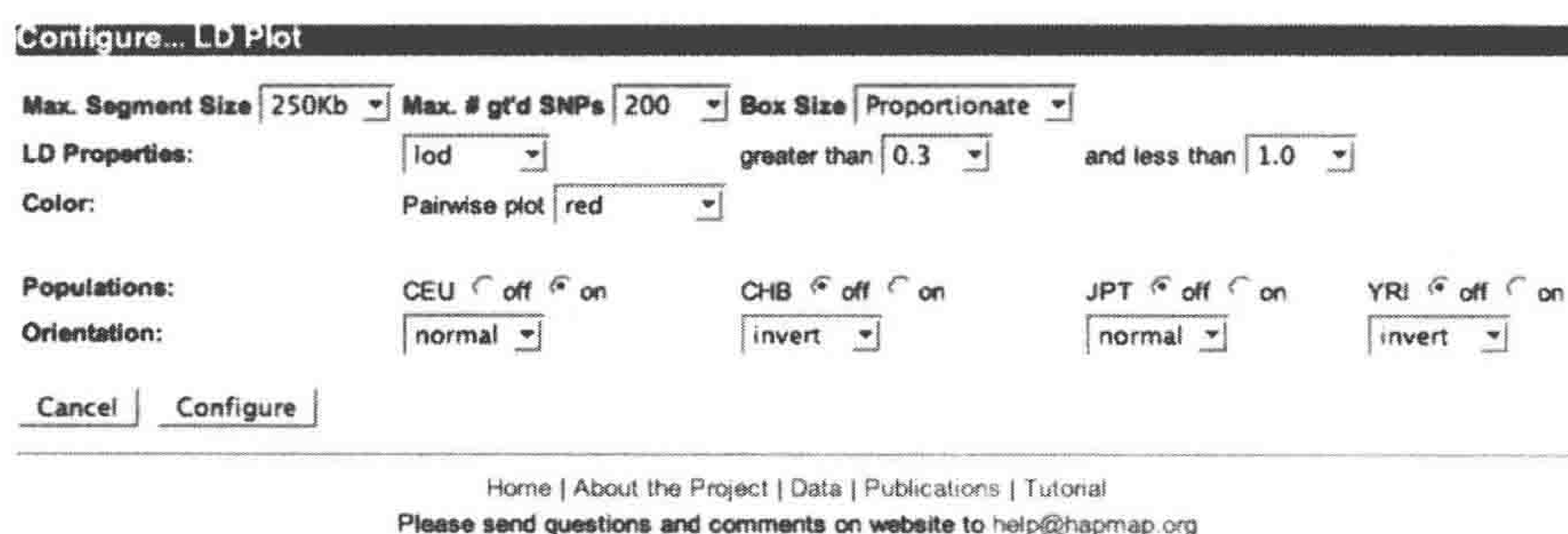



图 6-3 HapMap 基因组浏览器的图像页允许使用者选择不同风格的数据显示

(10) 点击“Configure”选项返回主要显示。该显示现在将出现对应每个所选择的群体的三角图（图 6-4）。

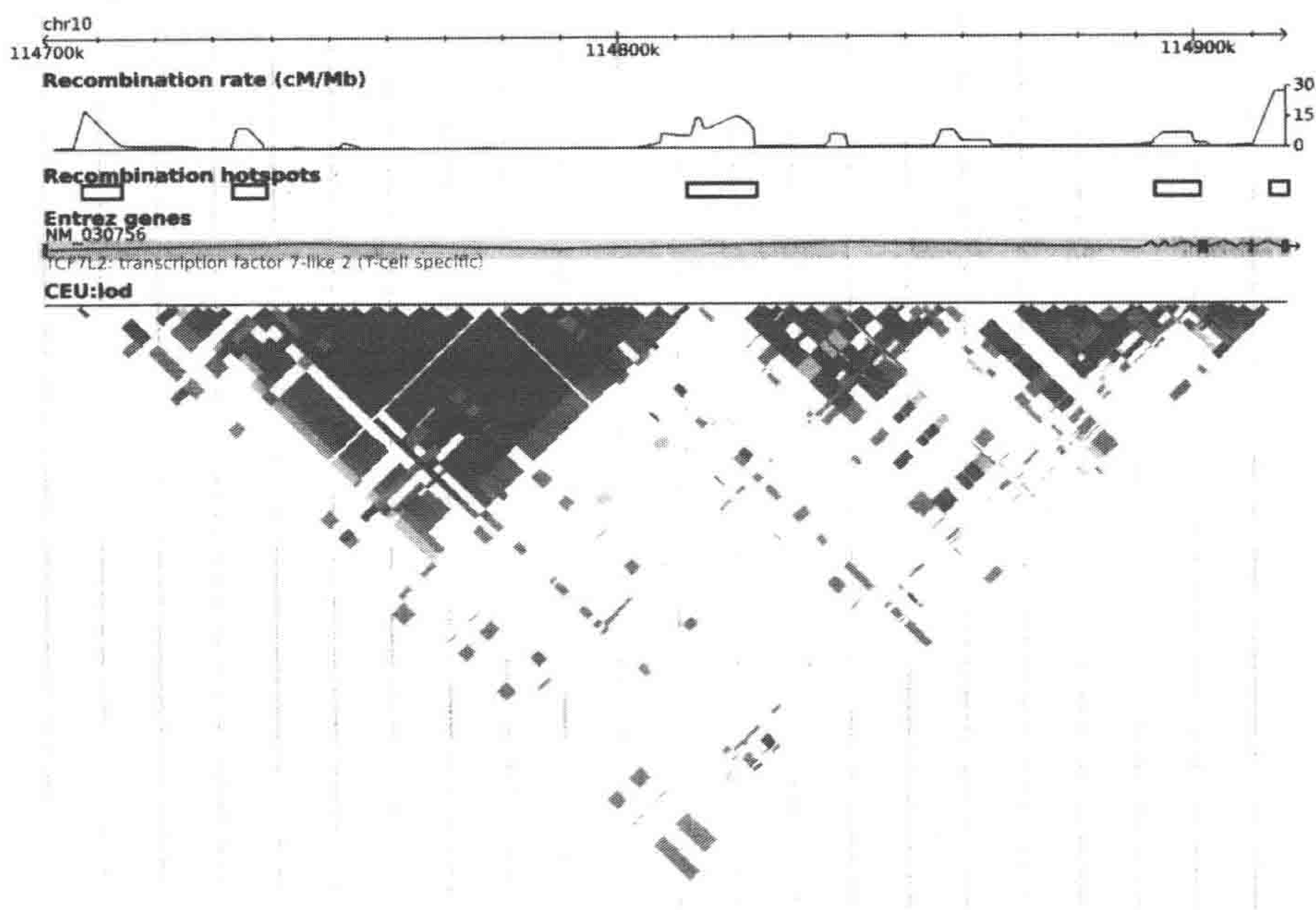



图 6-4　HapMap 基因组浏览器显示多重群体的 LD 值的三角图。

在有许多基因型 SNP 的区域，LD 插件会显著延长网页加载的时间。你可以在任何时候取消选定浏览器“Tracks”部分的相应检验盒关掉 LD 显示。LD 插件的设定储存在浏览器的一个 cookie 上，因此每次打开 LD 插件时没必要访问结构页。

(11) 传统的 $D'$ 和 $r^2$ 度量反映了两个 SNP 之间成对 LD 的程度，但在不同的大小度量时它们的敏感度和特异性不同。关于这些测量方法实际运用的讨论参见 Mueller (2004)。Daly 等 (2001) 描述了在 HapMap 网站显示所使用的 LOD 的度量。

图 6-4 是 HapMap 基因组浏览器显示多重群体的 LD 值的三角图。这里显示了由相对限定较好的低 LD 边界分开的有高 LD 斑点的典型 LD 区。三角图是由沿着 45°线连接的每一对 SNP 到水平条目线所构成。两个 SNP 相交位置的钻石颜色说明了 LD 的量：更深的颜色表示较高的 LD，一个灰色的钻石表示数据丢失。

## 提取和访问 tag-SNP

tag-SNP 是一套简化的 SNP，它集中了这些区域的许多 LD，可以被用来作关联研究，以减少为发现感兴趣的性状和基因组的一个区域之间相关性，在进行 LD 分析时所需的 SNP 的数量。对于较小的区域，可以使用图形和数字显示上面产生的 LD 以手动选择 tag-SNP，但为了最好的结果，研究者最好使用一种通过系列标签获取的最大数量的连锁 SNP 来选择 tag-SNP 的运算方法。不存在一套单一的 tag-SNP 能够满足每个关联研究设计的不同要求。研究者可能希望选择在特殊的基因型系统中工作良好的 SNP

(如那些已经被包含进特殊“SNP 芯片”的 SNP)，还可能希望找到在研究一个群体基因型时所需价格和他们能够发现的相关性范围之间的种种折衷方案。基于这样的理由，HapMap 网站没有提供一套事先选定的稳定的 tag-SNP，而是根据使用者提出的标准提供给研究者一个能够交互式选择 tag-SNP 的工具。Tagger 程序的运算法则产生了 tag-SNP 列表（http//www. broad. mit. edu/mpg/tagger/；de Bakker et al. 2005）。

（12）找到感兴趣的区域［步骤（1）～（4）］。

（13）在“Reports and Analysis”菜单中，选择“Annotate tag SNP Picker”选项。

（14）点击“Configure”以选取合适的 tag-SNP 选择（图 6-5）。选项包括。

- 选择群体和运算法则。
- 加载一个 SNP ID 的列表，将其包含进 tag-SNP 组中。
- 加载一个 SNP ID 的列表，使其排除在 tag-SNP 组之外。
- 加载一个对每一个 SNP 设计计分的列表（优先）。
- 选择对包含在该组中的 SNP 的最小可接受 LD 值和基因频率的限度。

（15）点击“Configure”钮以运行分析并返回主要显示中，结果显示在新的特征条目中（图 6-5）。正像如上的 LD 显示，设定被保存在浏览器的 cookie 中，不需要时可以关闭这个插件的条目。

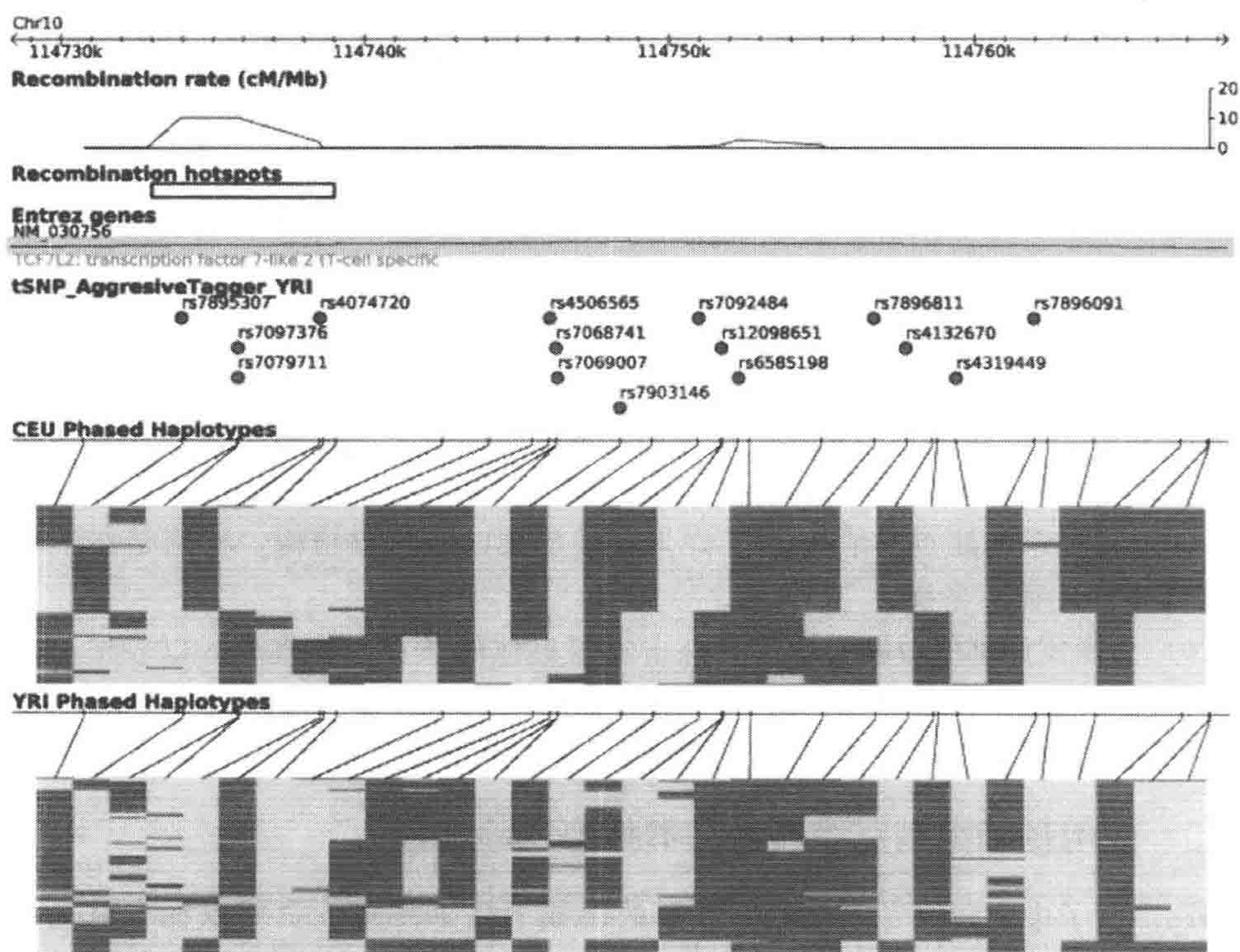



图 6-5　HapMap 基因组浏览器的 tag-SNP 和定相的单倍型的图形显示

## 浏览分相的单倍型

研究者可能希望将由 tag-SNP 选择运算法则选择出来的 tag-SNP 组与这一区域潜在的单倍型结构相联系。一种方法是同时打开配对的 LD 和 tag-SNP 条目（各自的步骤 (7) ～ (11) 和步骤 (12) ～ (15)。另一种方法是激活一个显示分相的单倍型的条目。在这一部分描述的分相的单倍型数据是 HapMap 计划协会使用 PHASE version 2.1 程序（Stephens and Donnelly 2003）产生的。在分相时，基因型中的每一个等位基因分配给两个亲代染色体之一；可通过最大拟然法对此进行运算，该运算在 HapMap 群体组中使用三重（世系）信息；如果三重信息不能使用，则可通过使数据与一个使历史上群体中潜在交叉数目最小化的模型相匹配来做到这一点。分相的单倍型作为一个图来显示，图中该计划采集来的个体的每一条染色体由一个高像素的线代表，每一个 SNP 等位被随机地染成蓝色或黄色。一个高 LD 的区域将被显示为有一长串 SNP 在多重的染色体共享等位基因，表示它们之间几乎没有重组。一个低 LD 区域将显示为排列较短和更加断续的区域。

(16) 进入一个感兴趣的区域 [步骤 (1) ～ (4)]。

(17) 从“Reports and Analysis”菜单中选择“Annotate Phased Haplotype Display”。

(18) 点击“Configure”设定单倍型显示的选项。这些选项使你能够选择显示单倍型信息的群体。

(19) 在选择了所需要的群体后点击“Configure”返回主要显示。每个所选择的群体将出现一个新的特征条目。每个条目用前面描述的双色方案显示该群体的单倍型（图 6-5）。

染色体的顺序由一个快速的分等级的聚类方法所决定，该方法将含有相似单倍型的染色体放在一起。

与配对 LD“三角显示”相比，这种显示的优点是它更加紧凑，因此比较适合显示大的区域。这使将长的共同单倍型的位置与被 tag-SNP 采集器选出的 SNP 相互联系变得容易。这种显示的缺点是它隐藏了许多该地区 LD 的细致结构，特别是在不相近的 SNP 之间强的连锁不平衡。

(20) 为恢复详细的分相基因型，点击需要群体的条目。这将进入以表格形式提供单倍型信息的网页。该表的每一行是一条单独的染色体，每一纵列是各自的 SNP。每个表项目的背景被设定成与在图形条目中所见的颜色相配的颜色。

## 方案二　使用基因组浏览器产生文本报告

在许多例子中，研究者愿意从感兴趣的区域下载 HapMap 资料以在本地进行整理分析。网点允许直接下载基因型、频率、tag-SNP 和来自基因组浏览器的其他报告。

# 方法

## 产生一个基因型列表文本

当研究者已浏览了图形区域并关注一个候选基因及其周围区域时，他可能想创建一个空间限定的文本储存处放置跨越该区域的基因型结果。然后这些资料进入一个 Excel 电子数据表或其他数据分析工具中。

（1）进入一个感兴趣的区域［方案一，步骤（1）～（4）］。

（2）从“Reports and Analysis”菜单中选择“Download SNP genotype data”。

（3）点击“Configure”。这将打开一个结构网页，允许你选择需要的 HapMap 群体，或者保存这些数据到磁盘中，或者在网页浏览器中浏览。

（4）选择需要的选项。点击“Go”，得到数据并产生一个报告。文本储存（text-dump）格式由一套含有每种 SNP 的 dbSNP ID 的行、两个对照的可替换的等位基因、基因组中 SNP 的位置和在选择的 HapMap 群体中每个个体 SNP 基因型组成。由于这种格式与用于批量下载的文件格式相同（方案五），它可以容易地载入研究者自己计算机上的 HaploView 程序（http://www.broad.mit.edu/mpg/haploview；Barrett et al. 2005）进行详细的分析。储存（dumper）结构设置储存在一个浏览器 cookie 上，下一次你可以点击主页上的“Go”钮并直接储存数据，不必首先设定 dumper。

## 产生基因型频率的文本列表

研究者可能希望下载感兴趣区域的等位基因频率的概要，然后研究者从中选择适合特定标准的 SNP，如那些特定群体中高度多态的 SNP。本方案描述创建一个标签限定的在特定基因组区域的 HapMap 基因频率资料的概要。

（5）进入一个感兴趣的区域［方案一，步骤（1）～（4）］。

（6）从“Reports and Analysis”菜单中选择“Download SNP Genotype Frequency Data”。

（7）点击“Configure”。

（8）点击“Go”得到数据并产生报告。由这一部分产生的报告包括每个由各自的 dbSNP ID 组成的各占一行的 SNP、它的基因组位置、在选择的群体中见到每种可能基因型的次数和此 SNP 在群体中的杂合性。

## 产生连锁不平衡值的文本列表

在选择了一个感兴趣的区域并实际检查了跨越感兴趣基因的 LD 的程度后，研究者可能希望下载由选定区域标签限定的该 LD 值的数字摘要。这一信息可用来选择一套“tag”（标签）SNP，这将作为其他伴随高度的 LD 的 SNP 的代理。

（9）进入一个感兴趣的区域［方案一，步骤（1）～（4）］。

（10）从“Reports and Analysis”菜单中选择“Download HapMap LD Data”。

（11）点击“Configure”。

(12) 点击“Go”得到数据并产生报告。由该方案产生的报告将显示在各自250 kb之内的所有SNP之间配对的LD。报告的每一行对应一对SNP。前两列说明SNP在染色体上的位置，第3列是已经计算出LD值的群体，第4列和第5列说明SNP对的db-SNP ID。随后是D′、$r^2$和两个SNP之间LD的LOD计分。

### 产生tag-SNP的文本列表

使用方案一，步骤(12)～(15)所描述的交互式tag-SNP选择条目，研究者可以调整选择标准直到对标签组的特点感到满意。本方案描述如何在此产生一个SNP的文本积存(dump)，以便在与其他HapMap产生的报告信息结合时被用来创建一个筛选组。

(13) 进入一个感兴趣的区域[方案一，步骤(1)～(4)]。

(14) 从“Reports and Analysis”菜单中选择“Download tag SNP Data”。

(15) 点击“Configure”为tag-SNP选择设立交互式选项，像方案一步骤(12)～(15)所描述的那样。选项包括

- 选择一个群体和一种运算法则。
- 上载一个SNP列表在必要时作为tag-SNP包含进去。
- 上载一个排除在tag-SNP列表之外的SNP列表。
- 设定终止LD和等位基因频率的值。

(16) 点击“Go”得到数据并产生报告。所产生的报告包含一个tag限定的tag-SNP名称列表、染色体、位置和此区域的基因频率。随后罗列tag-SNP、它们获得的非tag-SNP和每个tag-SNP与它们获得的非tag-SNP之间LD的强度。

## 方案三 用HaploView操作HapMap数据

希望对高LD区域的显示进行精细调控，或希望对tag-SNP试验新的算法的高级使用者可能希望使用HaploView程序(Barrett et al. 2005)分析HapMap数据。这一程序可以与HapMap基因组浏览器很好地结合。HaploView对比HapMap网站基因组浏览器的一个很大的优点是它同时显示LD区域的高和低能力的观察，并能在卷轴和缩放过程中给予即时反馈。HaploView的其他优点可以在第21章找到。

## 方法

(1) 安装HaploView，登录http://www.broad.mit.edu/mpg/haploview/，点击下载链接。

HaploView要求在本地机上安装Java Runtime Environment (JRE)，如果未安装，可以在http://www.java.com上找到最新版本。

(2) 下载适合你计算机的HaploView程序。

i. 对于Windows，下载Windows安装文件。双击安装文件在开始菜单创建一个HaploView文件夹。

ii. 对于 MacOS X and Unix，下载 HaploView. jar 文件。

(3) 参照方案二，步骤 (1) ～ (4)，从感兴趣的区域下载基因型。

(4) 打开 HaploView. jar 文件开始 HaploView。

i. 对于 Windows，从开始菜单的 HaploView 文件夹打开 Haploview. jar。

ii. 对于其他操作系统，双击 HaploView. jar 文件。

(5) 在 HaploView 的欢迎窗口，点击“Load HapMap Data”装载基因型。

(6) 浏览、下载并打开含有基因型的文件。

一旦数据被装载，HaploView 将提供给你一些选择，查看跨越该区域的高分辨 LD 三角图，查看分享的单倍型和它们的重组频率，用几种方法选择 tag-SNP 组。

(7) 在 HaploView 窗口的上面选择合适的指定标签在这些分析和显示中进行选择。

(8) 此外，可以从 HapMap 站点直接将基因组条目加入 HaploView 显示中，包括 Entrez 基因条目和重组率。

(9)“Display”和“Analysis”菜单可以让你改变 LD 三角图的大小和颜色，也可在众多定义单倍型构件的运算法则和高相互 LD 的 SNP 区域之间进行选择。

## 方案四 使用 HapMart 得到 HapMap 数据

基于运行的考虑，通过基因组浏览器进行 HapMap 数据交互式访问的接口被限制在不大于 5 Mb 宽度的区域内。希望获得染色体范围或基因组范围数据的研究者可以有两个选择：批量下载或通过 HapMart 进入。本处提到的 HapMart 允许研究者利用不同的标准选择 SNP 并只显示那些他们感兴趣方面的资料。

## 方法

(1) 在 http://www.hapmap.org/BioMart/martview 中进入“MartView”界面。点击“next”使用缺省数据库和数据包开始一个新的询问。

(2) 在过滤器页面上 (filter page) (图 6-6) 根据众多的标准选择需要的 SNP (各自的或结合的)。你可以以任何顺序应用过滤器并通过“next”和“back”修订现有的过滤器。当你使用过滤器时，所选择的 SNP 的数目被显示在右侧摘要栏。可用的过滤器包括

- 获排除的 SNP 列表。
- 所选的 SNP 低基因频率的最小值。
- 将 SNP 限制在内含子区、mRNA/UTR 区、编码的非同义区或编码的同义区。
- 将 SNP 限制在特定的基因组区。
- 将 SNP 限制在与特定的基因 ID 交叠的区域。

(3) 在选择和细化合适的过滤器后，点击“next”，进入输出选项页。该页允许你选择希望输出到报告中的领域。有许多被设计成一系列顺着屏幕上端排列的标签一样的输出选项。为得到感兴趣的信息，选择页面顶端适当的标签，然后选择合适的检验盒。

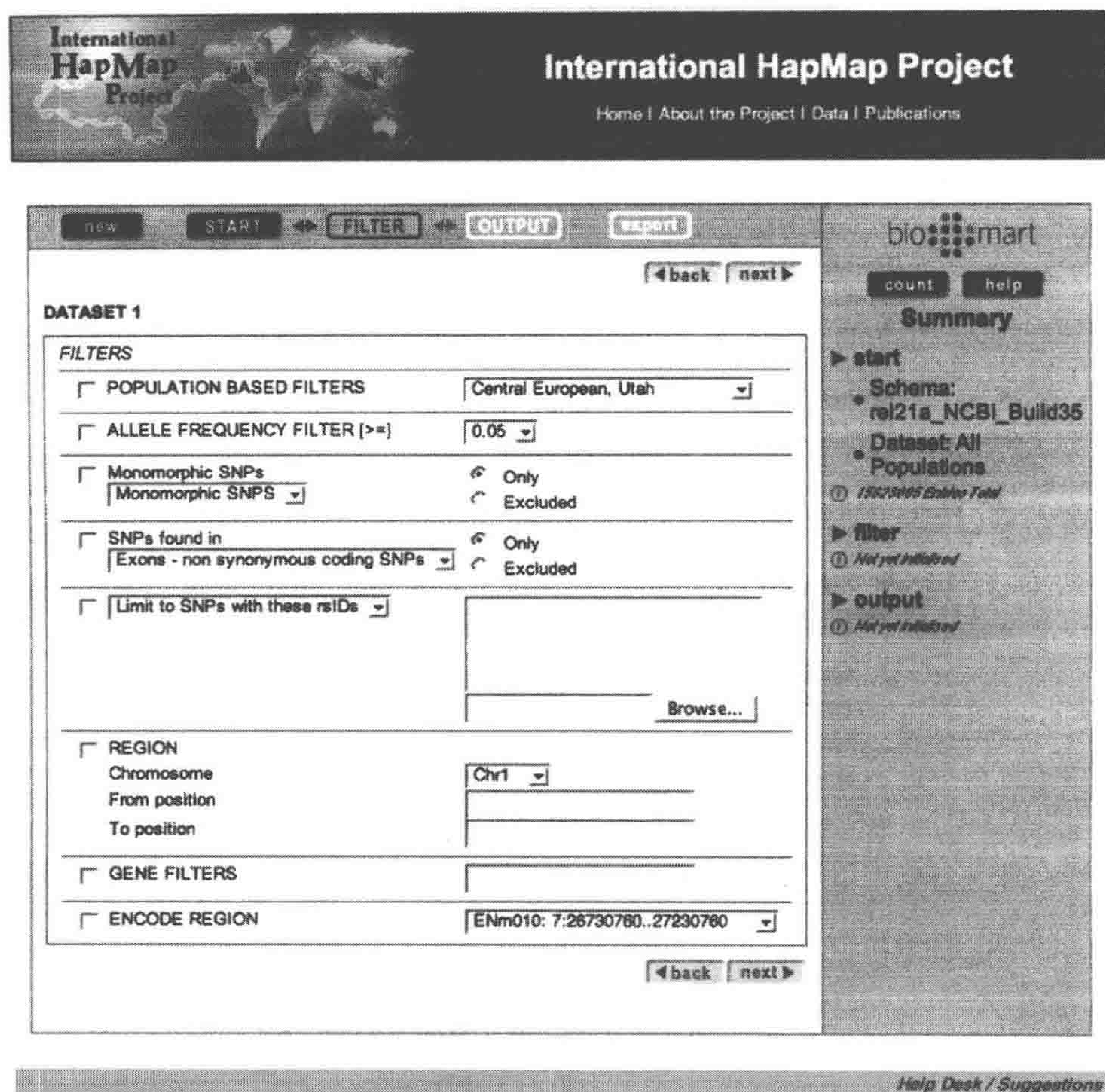



图 6-6　HapMart 界面适合于根据特定标准过滤 HapMap 数据

输出选项包括

- 基因型。
- SNP 染色体位置。
- 等位基因。
- 基因型频率。
- 基因频率。

(4) 如果得到的 SNP 数目很大，你可以选择“gzip file compression”，让这些报告在进入浏览器之前被压缩，从而减少下载报告的时间。

(5) 为得到结果，选择“export”。

由 HapMart 产生的报告是标签限定的文本格式，适合输入 Excel 或相关的数据库。HapMart 背后的引擎是一个称为 BioMart 的普通数据挖掘结构。

## 方案五 通过批量下载取得数据

染色体范围或基因组范围数据的批量下载提供了全 HapMap 数据包文本积存。然而，这种下载没有提供任何过滤或选择服务。

## 方法

（1）为得到批量数据，浏览下载页（http://www.hapmap.org/downloads/；下载储藏库可通过在 ftp://www.hapmap.org 的匿名 FTP 上得到）。数据文件被染色体/区域和群体所分割，分 3 个种类得到，每种有自己的子目录。

- non-redundant/：这些数据包对每个 SNP 群体只包括一套基因型。所有的基因型包已经通过质量控制检查，同一个 SNP 的多重递交（由于 QA 操作、校正的递交、项目的计划冗余所产生）已被排除。这是多数使用者所希望的数据包。
- redundant-filtered/：所有这些数据包都通过质量控制检查，但冗余数据没有被排除。
- redundant-unfiltered/：该数据包包含所有由不考虑质量控制检查的计划作分型的 SNP 基因型。这些“未加工的”数据适合于希望看到具有潜在生物学意义的数据的使用者，这些数据通常被项目质量控制检查过滤掉了。

（2）从可利用的批量数据下载列表选择合适的链接。

i. 点击“Genotypes”链接到基因型下载目录。“latest/”子目录总是指向现有的数据冻结。

ii. 点击“LD Data”链接到 LD 数据下载目录。“latest/”子目录再次指向最近可利用的数据冻结。LD 值被表示为 $D'$、LOD 和 $r^2$ 值。

iii. 为下载分项的基因型，点击“Phasing Data”链接。这将带你到一个含有表示 PHASE 程序输出的数据文件的目录。

（3）除了以上列出的批量数据包之外，还有一系列其他批量下载可以利用。计算的频率值可用于该计划所分型的 SNP，还有来自一些基因型的原始信号强度数据（目前仅来源于 affymetrix genechip 100k and 500k mapping arrays）。可以下载重组率数据，像该计划中使用的方案所做的，该计划中基因型分型的样品的细节也可下载。

## 讨论

已经发展起一系列公共在线资源作为高通量全基因组数据包的入口。UCSC 基因组浏览器（http://genome.UCSC.edu；Kent et al. 2002）和 EnsEMBL 计划（http://www.ensembl.org；Hubbard 2007）已经发展出多物种的基因组浏览器，显示图形基因组注释并提供潜在的数据。dbSNP（Wheele et al. 2007）是一个有关单核苷酸多态信息的知识库，但还没有广泛的关于这些 SNP 相互关系的信息。

位于 http://www.hapmap.org 的 HapMap 网站有一个清楚的焦点。它的目的是作

为一个资源，对高批量、高质量、全基因组范围内的人类遗传资料进行显示、获取和分析，并强调为推进疾病相关研究在工具上提供支持。尽管这一资源仍在发展中，目前它提供 HapMap 计划所调查的群体中常见多态的可视类型的基本工具，根据多种标准选择 tag-SNP 组，产生精选的用户化数据组。将来 HapMap 网站将发展到能为设计和解释遗传相关研究提供更多的服务。

## 参考文献


Barrett J.C., Fry B., Maller J., and Daly M.J. 2005. HaploView: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

Conrad D.F., Andrews T.D., Carter N.P., Hurles M.E., and Pritchard J.K. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38:** 75–81.

Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J., and Lander E.S. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29:** 229–232.

de Bakker P.I.W., Yelensky R., Pe'er I., Gabriel S.B., Daly M.J., and Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* **37:** 1217–1223.

Hinds D.A., Kloek A.P., Jen M., Chen X., and Frazer K.A. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38:** 82–85.

Hubbard T.J.P., Aken B.L., Beal K., Ballester B., Caccamo M., Chen Y., Clarke L., Coates G., Cunningham F., Cutts T., et al. Ensembl 2007. 2007. *Nucleic Acids Res.* **35:** D610–D617.

Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., Scherer S.W., and Lee C. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* **36:** 949–951.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., and Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res.* **12:** 996–1006.

McCarroll S.A., Hadnott T.N., Perry G.H., Sabeti P.C., Zody M.C., Barrett J.C., Dallaire S., Gabriel S.B., Lee C., Daly M.J., and Altshuler D.M. 2006. Common deletion polymorphisms in the human genome. *Nat. Genet.* **38:** 86–92.

Mueller J.C. 2004. Linkage disequilibrium for different scales and applications. *Brief Bioinform.* **5:** 355–364.

Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shapero M.H., Carson A.R., Chen W., et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.

Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Maner S., Massa H., Walker M., Chi M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Sharp A.J., Locke D.P., McGrath S.D., Cheng Z., Bailey J.A., Vallente R.U., Pertz L.M., Clark R.A., Schwartz S., Segraves R., et al. 2005. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77:** 78–88.

Smith A.V., Thomas D.J., Munro H.M., and Abecasis G.R. 2005. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.* **15:** 1519–1534.

Stephens M. and Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73:** 1162–1169.

Thorisson G.A., Smith A.V., Krishnan L., and Stein L.D. 2005. The International HapMap Project Web site. *Genome Res.* **15:** 1592–1593.

Tuzun E., Sharp A.J., Bailey J.A., Kaul R., Morrison V.A., Pertz L.M., Haugen E., Hayden H., Albertson D., Pinkel D., Olson M.V., and Eichler E.E. 2005. Fine-scale structural variation of the human genome. *Nat. Genet.* **37:** 727–732.

Vastrik I., D'Eustachio P., Schmidt E., Joshi-Tope G., Gopinath G., Croft D., de Bono B., Gillespie M., Jassal B., Lewis S., et al. 2007. Reactome: A knowledge base of biologic pathways and processes. *Genome Biol.* **8:** R39.

Wheeler D.L., Barrett T., Benson D.A., Bryant S.H., Canese K., Chetvernin V., Church D.M., DiCuccio M., Edgar R., Federhen S., et al. 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35:** D5–D12.


## 互联网资源

**http://www.broad.mit.edu/mpg/haploview/** HaploView bundles many everyday analysis tasks into one easy-to-use package. It is an open-source program written in Java and capable of running on Windows, MacOS, and UNIX platforms.

**http://www.ensembl.org** Ensembl produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl concentrates on vertebrate genomes, but other groups have adapted the system for use with plant and fungal genomes.

**http://www.hapmap.org** The primary portal to genotype data produced as part of the International Haplotype Map Project. The HapMap Web site provides researchers with a number of tools that allow them to analyze the data as well as to download data for local analyses.

**http://www.java.com** This site has downloads of the Java Runtime Environment available for Windows, Macintosh, and UNIX.

**http://www.ncbi.nlm.nih.gov/SNP** dbSNP is the central standard repository for genotyping information, including HapMap as well as other targeted resequencing projects.

**http://www.reactome.org** Reactome is a free on-line curated resource of core pathways and reactions in human biology. In addition to curated human events, inferred orthologous events in 22 non-human species including mouse, rat, chicken, zebra fish, worm, fly, yeast, two plants, and *E. coli* are also available.

**http://www.broad.mit.edu/mpg/tagger** Tagger can be used for tag-SNP selection and evaluation using HapMap data.

# 7 植物DNA的分离及其基因型分析

Nathan M. Springer

*Department of Plant Biology, University of Minnesota, St. Paul, Minnesota 55108*



## 简介

分离植物DNA有多种方案（Saghai-Maroof et al. 1984；Dellaporta et al. 1985；Doyle and Doyle 1987；Dard and Dronavalli 1992；Wang et al. 1993；Richards et al. 1994；Csaikl et al. 1998），其中一个常见的就是存在细胞壁。为了有效地分离植物DNA，用物理或酶解的方法来降解细胞壁是很有意义的。另外，一些组织（如胚乳）和富含淀粉或苯酚复合物的特殊品种，它们的DNA分离会更复杂。很多植物DNA分离方案都是针对特定品种的难题而设计的（Baker et al. 1990；Weeb and knapp 1990；Maliyakal 1992；Lohdi et al. 1994；Porebski et al. 1997）。

分离植物DNA的方案可以被分为三组。第一组方案能够得到大量DNA用于Southern印迹法（Chen and Dellaporta 1994）。通常，这些方案要有5～15 mL的提取体积，并且要求用研钵、杵或匀浆机将组织磨碎。第二组方案能分离较少量的DNA用于PCR（见下面的方案或Qiagen的植物DNA小量提取试剂盒）。一般情况下，这类方案也要求有研钵和杵的研磨或者研磨机、染色摇匀器（paint shaker）的混匀。它们通常消耗中等量的DNA，这些DNA要求能够长时间地稳定保存。第三组方案利用商业试剂一步快速地分离DNA（见Manen et al. 2005或Sigma的N-Amp植物PCR DNA提取试剂盒）。多数情况下一种提取试剂加入植物样品中，样品经简单的加热就会释放出DNA。尽管这些方案经常是最快速的，但是分离得到的DNA不能稳定地储存，而且有可能被植物中的一些复合物污染，这些复合物还会阻碍PCR的进行。在网上搜索能克服特殊物种或组织分离DNA障碍的植物DNA分离方法，网络会提供多种相应方案。除此之外，因特网上的搜索还能够鉴别一些用于植物DNA分离的商业产品，包括来自Sigma、Qiagen、Epicentre、Invitrogen、MO BIO、Clontech、Whatman、Promega、Bio-Nobile的产品，他们通常提供其最适宜的组织和品种。个人研究者需要平

衡花费和另外一些因素，如 DNA 的数量、DNA 的质量和现有的仪器的要求来决定最佳方案。

## 方案一　用于 PCR 的植物 DNA 分离以及用有机提取剂和 CTAB 进行基因型分析

Nathan M. Springer

*Department of Plant Biology, University of Minnesota, St. Paul, Minnesota 55108*

以下是一个相对简单的方案，这个方案适用于许多植物种。它可提供用于 PCR 的大量的高质量且能长时间保持稳定的 DNA，每个样品的平均花费也很低。另外，这个方案相对简易，相关训练较少的人也可以操作。一个普通的大学生用这个方案每天能够进行 200～300 个样品的分离。这个方案的不足是它要用到一个冰冻干燥器和一个研磨机或者一个类似于染色摇匀器（paint shaker）的设备，其中有一步要用到一种有机提取剂，还需要用到通风橱。

## 材料

注意：带有＜!＞标识的试剂要小心操作，具体见附录。

### 试剂

氯仿＜!＞[1]（Fisher）：异戊醇＜!＞（Fisher）（24：1）

CTAB 提取缓冲液：

100 mL 1 mol/L 氨基丁三醇（pH 7.5）（Fisher）、140 mL 5 mol/L NaCl（Fisher）、20 mL 0.5 mol/L EDTA（Fisher）溶有 10 g 的 CTAB＜!＞（十六烷基三甲基溴化铵）（Calbio Chem）。混合后加水至 990 mL。在需要使用前再加 10 mL β-硫基乙醇＜!＞，摇晃混匀。最终的溶液含有 0.1 mol/L 氨基丁三醇（Fisher）、0.75 mol/L NaCl、0.01 mol/L EDTA、1% CTAB、1% β-硫基乙醇。

乙醇（70%）

异丙醇＜!＞（Fisher）

植物组织样品（如 1～2$cm^2$ 玉米幼苗的叶子）：通常，较年幼的组织能够提供最好质量和最多数量的 DNA。此方案也已经用于根组织、丝组织和成熟叶片。

氨基丁三醇 Tris（1 mmol/L，pH 8.0）（Fisher）

### 仪器

冰冻干燥器

玻璃珠（5 mm；Fisher）

Retsch 300 细胞粉碎仪（Qiagen）：可以用修改过的染色摇匀器或竖锯替代。

试管（1.5 mL，2 mL）

移液管或 1 mL 移液器

漩涡式搅拌机

水浴锅（65℃）

## 方法

（1）收集植物组织到一个 2 mL 的试管中。

（2）冰冻干燥 1～2 d。干燥时间决定于组织类型和冰冻干燥仪的干燥能力。一旦组织变得易碎就可以研磨了。

（3）向每支试管加入两个 5 mm 的玻璃珠并用 Retsch 300 细胞粉碎仪在 30 r/min、2 min 的条件下粉碎组织。每套 24 份样品要求研磨 5 min。

（4）向每支试管中加入 700 μL CTAB 提取缓冲液并倒转混匀。

（5）65℃下保温 30～60 min，每 15 min 倒转混匀一次。

（6）加入 400 μL 24∶1 氯仿∶异戊醇并用漩涡式搅拌机混匀。

（7）10 000 *g* 条件下离心 5 min。

（8）用移液管或 1 mL 的移液器将上述水相转移到新的 1.5 mL 试管中。

可选择项：如果要进一步纯化 DNA 以长时期储存，重复步骤（6）～（8）。

（9）加入 300 μL 异丙醇并倒转混匀。

（10）10 000*g* 条件下离心 5 min。可见发白的 DNA 小球。

（11）将液体倒掉，加入 500 μL 70%乙醇并用漩涡式搅拌机混匀。

（12）10 000*g* 条件下离心 2 min 然后倒掉上层液体。

（13）干燥所得 DNA 小球，然后用 250 mL 1 mmol/L 的 Tris（pH 8.0）溶液重新溶解（见疑难解答）。每 24 份样品的提取［步骤（3）～（13）］要求约 1.5 h，但是多组（每组 24 份）样品可以同时进行提取。

## 疑难解答

问题［步骤（13）］：DNA 小球不能重新溶解。

解答：这可能是被淀粉污染的结果。用较少量的组织来制备或用专门为淀粉含量高的组织设计的方案来进行 DNA 分离。

问题［步骤（13）］：分离得到的 DNA 很少或没有。

解答：试一下更幼嫩的组织，在提取缓冲液中保持更长的保温时间，或者换一种方案查阅资料来确定是否有我们所研究物种适合的特殊方案。

问题：DNA 碎片过多。

解答：这可能是由于环境胁迫或者植物组织的处理不当造成的。将植物组织放在冰块上或者放在冰冻干燥仪中 0.5 h 可以减少碎片。

问题：DNA 提取成功但是 PCR 失败了。

解答：这可能是因为氯仿的污染或者是存在某个物种的特异性化合物抑制了 PCR

的反应。尝试使用分离柱方案，如 Qiagen DNeasy 植物试剂盒。

## 方案二　运用 96 孔平板和 CTAB 从冰冻干燥处理的植物组织中提取 DNA

An-Ping Hsia,[1] Hsin D. Chen,[1] Kazuhiro Ohtsu,[1] and Patrick S. Schnable[1,2,3]

*[1]Department of Agronomy, [2]Department of Genetics, Development, and Cell Biology, and [3]Center for Plant Genomics, Iowa State University, Ames, Iowa 50011*

这个方案是 DNA 分离方法（Dietrich et al.　2002）的一个改进版本，运用到 CTAB（Rogers and Blendich　1985）和 96 孔平板。该方案产量高，有利于大量种群的图谱分析。它已经被应用于玉米和拟南芥的叶组织的 DNA 提取中，而且从中提取的 DNA 量足够做 100～500 次 PCR。

## 材料

注意：带有<! >标识的试剂要小心操作，具体见附录。

### 试剂

氯仿（～0.75% 乙醇作为保存，Technical grade，Fisher Scientific）<!>/辛醇 <!>（24∶1）

CTAB 提取缓冲液

每升：

| [原料] | [添加量] | [最终浓度] |
| --- | --- | --- |
| 1 mol/L 氨基丁三醇，pH 7.5 | 100 mL | 100 mmol/L |
| 5 mol/L NaCl | 140 mL | 0.7 mol/L |
| 0.5 mol/L EDTA | 20 mL | 10 mol/L |
| * BME（β-巯基乙醇，14 mol/L stock）<!> | 10 mL | 1% |
| CTAB（Sigma）<!> | 10 g | 1% |
| $H_2O$ | 730 mL | — |

* 为了得到最佳结果请在使用前再加入 BME。

乙醇（70%；Aaper 酒精）

异丙醇 <!>

TE [可选择的，见步骤（17）]

来自幼苗（种植 5～7d）的幼嫩叶片的组织样品

### 仪器

结合板（Qiagen）

气孔板（Qiagen）

夹具

收集架（96孔）

冰冻干燥器

玻璃珠（1.7～2.5 mm，MO-Sci Corporation）

恒温箱（58℃）

染色摇匀器（红魔设备 Red Devil Equipment）

纸巾（厚，Kim Towels，Erie Cotton 产品）

PCR 平板（96孔，skirted）

移液管（200 μL）

胶质玻璃板（10.75 in 长［L］×8.5 in 宽［W］×3/8 in 深［D］）（1 in＝2.54 cm，后同）

分离管（1.2 mL，带盖）

V底盘（96孔，0.6 mL，Costar/Fisher）

木板（11.5 in L×4.5 in W×0.75 in D）

## 方法

（1）在开始收集前按下述步骤预先向1.2 mL分离管（在96孔架上）中加入1.7～2.5 mm玻璃珠：

i. 向96孔的PCR板（0.2 mL管）中填入1.7～2.5 mm的玻璃珠。

ii. 用200 μL移液管将卡在小管之间的玻璃珠移除。

iii. 小心地将装有玻璃珠的平板在96孔的收集架上倒置，保证平均每个管中玻璃珠等体积。

（2）将组织样品放入1.2 mL的分离管中。

（3）用气孔板将每个PCR板的小管密封并冰冻干燥1～3 d。

（4）去掉气孔板并用分离管帽代替。在两块木板（11.5″L×4.5″W×0.75″D）中间用两个收集板加固。将它们放入一个染色混匀器中，用染色混匀器将冰冻干燥过的组织样品粉碎5～10 min。最终每份样品都应为均匀的粉末状。

（5）缓慢地取出收集架并快速离心以固定粉末。小心地移开分离管帽加入600 μL CTAB提取缓冲液，盖上分离管帽并缓缓倒置混匀。

（6）58℃恒温保持15 min。反复颠倒样品约10 min，再在58℃下恒温保持15 min。

（7）将样品从恒温箱中取出并在通风橱中冷却10 min。

（8）2500 r/min下快速离心。

（9）加入300 μL 24：1的氯仿：辛醇。将板夹在Plexiglas（10.75″L×8.5″W ×3/8″D）和夹具之间。缓缓倒置5 min。

（10）3500 r/min下离心10 min。

（11）在步骤（9）中向0.6 mLV形底96孔平板每个孔中加入300 μL异丙醇。

（12）离心后，将250～300 μL上层液体转移到胶质玻璃板中，用结合板盖严并倒置10 min。

(13) 3500 r/min 下离心 10 min 以收集 DNA。

(14) 移去结合板，在水池上倒置胶质玻璃板，甩掉每个板上的液体。此时 DNA 小球应该在每个小管的锥形底部可见。

(15) 向每个小管中加入 50 μL 70%的乙醇洗 DNA 小球。3500 r/min 下离心3～5 min。

(16) 缓慢倒置板到厚纸巾上使乙醇被吸干，并使其干燥 10～15 min。

(17) 用 50～100 μL TE 或 $H_2O$ 使 DNA 小球再次溶解。DNA 溶解时间可以在 20 min直至过夜。DNA 溶液在－20℃（长期）或 4℃（短期）下储存。

## 方案三　运用 96 孔平板和基因纯化 DNA 提纯试剂盒来纯化拟南芥 DNA

Yan Li

*Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637*

这个方案描述了怎样运用适用于 96 平板的基因纯化 DNA 提纯试剂盒来高产量地提纯拟南芥 DNA。

## 材料

注意：带有<!>标识的试剂要小心操作，具体见附录。

### 试剂

拟南芥叶片组织（两整片叶片，～100 mg 鲜重；或者是整片幼嫩莲座叶）

乙醇（70%）

异丙醇（100%）<!>

基因纯化 DNA 提纯试剂盒（Gentra Systems），包含下列试剂：细胞裂解液、DNA 水合溶液、蛋白质沉淀溶液、RNase A 溶液

### 仪器

配珠器

桌面纸张（吸收剂）

Geno/Grinder（SPEX CertiPrep）

磨球（5 mm）

水浴锅（65℃）

冰

恒温箱（37℃）

平板（96 深孔）

橡皮垫

漩涡混合器

## 方法

(1) 收集叶片组织到 96 深孔的平板中（使用新鲜的或使用前一直在 −80℃ 条件下保存的叶片）。

(2) 用配珠器加一研磨球（5 mm）到每个孔中。

(3) 加入 100 μL 细胞裂解液。用橡皮垫盖上平板并封闭紧。

(4) 将平板放入 Geno/Grinder 研磨机中，在 400/min 冲下研磨 45 s。

(5) 移去平板并在 3000*g* 下离心 1 min。

相对离心力 RCF（*g*）= 1.12 r [r/（min · 1000）]$^2$，其中 r 为旋转半径，以 mm 计算。

(6) 加 280 μL 细胞裂解液到平板的每个孔中。

(7) 密封平板并将其置于 Geno/Grinder 研磨机中，在 400/min 冲下研磨 45 s，此时会有水沫出现。为了防止污染一定要封闭紧。

(8) 3000 *g* 下离心 1 min，高速漩涡旋转平板。

(9) 将平板放入在 65℃水浴锅中恒温保持 60 min。(保证封闭严密) 冷却至室温。

(10)（可选择项）加 3 μL RNase A 溶液到每个孔中，保证盖严后低速漩涡旋转。在 37℃恒温箱中保持 30 min（15～60 min）。

(11) 将平板放在冰上 1 min 以冷却至室温。

(12) 向每个孔中加入 133 μL 蛋白质沉淀溶液。

(13) 高速漩涡旋转 20 s 或者用 Geno/Grinder 研磨机研磨 20 s。

(14) 将平板放在冰上冷却 15～60 min。

(15) 4500 *g* 离心 15 min 在 10℃条件下收集蛋白质，如果这时的沉淀的小球不够紧密可以重复步骤（13）～（15）。

(16) 在离心时，用移液器吸取 200 μL 100%异丙醇到另一个平板中。

(17) 转移 200 μL 上层含有 DNA 的液体到新的含有 200 μL 100%异丙醇的平板中。

(18) 用新的盖子封闭平板并缓缓颠倒混匀 50 次。

(19) 4500*g* 离心 5 min。沉淀的 DNA 以小球的形式可见。

(20) 将整个 96 孔平板倒置在干净的吸收性桌面纸张上倒出上层残余的液体。重复 10 次进行干燥（注意小球）。

(21) 加入 200 μL 70%的乙醇并缓缓倒置 10 次洗 DNA 小球。

(22) 4500 *g* 离心 5 min。再将整个 96 孔平板倒置在干净的吸收性的桌面纸张上倒出上层悬浮的液体。倒置反复 10 次进行干燥（注意小球）。

(23) 将平板倒置 5～10 min 风干样品。

(24) 向每个孔中加入 50～75 μL DNA 水合溶液并密封平板 [可用双蒸水($ddH_2O$) 代替 DNA 水合溶液]。

(25) 室温下水合 DNA 过夜。

(26) DNA 溶液置于 4℃或 −20℃短期储存，−80℃长期储存。

# 参考文献


Baker S.S., Rugh C.L., and Kamalay J. 1990. RNA and DNA isolation from recalcitrant plant tissues. *Biotechniques* **9:** 268–272.

Chen J. and Dellaporta S. 1994. Urea-based plant DNA miniprep. In *The maize handbook* (ed. M. Freeling and V. Walbot), pp. 526–527. Springer-Verlag, New York.

Csaikl U.M., Bastian H., Brettschneider R., Gauch S., Meir A., Schauerte M., Scholz F., Sperisen C., Vornam B., and Ziegenhagen B. 1998. Comparative analysis of different DNA extraction protocols: A fast, universal maxi-preparation of high quality plant DNA for genetic evaluation and phylogenetic studies. *Plant Mol. Biol. Rep.* **16:** 69–86.

Dellaporta S.L., Wood J., and Hicks J.B. 1985. Maize DNA miniprep. In *Molecular biology of plants: A laboratory course manual* (ed. R. Malmberg et al.), pp. 36–37. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Dietrich C.R., Cui F., Packila M.L., Li J., Ashlock D.A., Nikolau B.J., and Schnable P.S. 2002. Maize *Mu* transposons are targeted to the 5′ untranslated region of the gl8 gene and sequences flanking *Mu* target-site duplications exhibit non-random nucleotide composition throughout the genome. *Genetics* **160:** 697–716.

Doyle J.J. and Doyle J.L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19:** 11–15.

Lodhi M.A., Ye G.-N., Weeden N.F., and Reisch B.I. 1994. A simple and efficient method for DNA extractions from grapevine cultivars and *Vitis* species. *Plant Mol. Biol. Rep.* **12:** 6–13.

Maliyakal E.J. 1992. An efficient method for isolation of RNA and DNA from plants containing polyphenolics. *Nucleic Acids Res.* **20:** 2381.

Manen J.F., Sinitsyna O., Aeschbach L., Markov A.V., and Sinitsyn A. 2005. A fully automatable enzymatic method for DNA extraction from plant tissues. *BMC Plant Biol.* **5:** 23.

Oard J.H. and Dronavalli S. 1992. Rapid isolation of rice and maize DNA for analysis by random-primer PCR. *Plant Mol. Biol. Rept.* **10:** 236–241.

Porebski S., Bailey L.G., and Baum B.R. 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* **15:** 8–15.

Richards E., Reichardt M., and Rogers S. 1994. Preparation of genomic DNA from plant tissue. In *Current protocols in molecular biology* (ed. F.M. Ausubel et al.), vol. 1, pp. 2.3.1–2.3.7. Wiley, New York.

Rogers S.O. and Blendich A.J. 1985. Extraction of DNA from milligram amounts of fresh herbarium and mummified plant tissues. *Plant Mol. Biol.* **5:** 69–76.

Saghai-Maroof M.A., Soliman K.M., Jorgensen R.A., and Allard R.W. 1984. Ribosomal DNA spacer-length polymorphism in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci.* **81:** 8014–8019.

Wang H., Qi M., and Cutler A. 1993. A simple method of preparing plant samples for PCR. *Nucleic Acids Res.* **21:** 4153–4154.

Weeb D.M. and Knapp S.J. 1990. DNA extraction from a previously recalcitrant plant genus. *Plant Mol. Biol. Rep.* **8:** 180–185.

# 8 从植物组织中制备 RNA

An-Ping Hsia,[1] Hsin D. Chen,[1] Kazuhiro Ohtsu,[1] and Patrick S. Schnable[1,2,3]

*[1]Department of Agronomy, [2]Department of Genetics, Development, and Cell Biology, and [3]Center for Plant Genomics, Iowa State University, Ames, Iowa 50011*



## 简介

RNA 在基因型分析中的应用对于特定的研究是有利的，因为转录物组显著地少于基因组并含有较少的重复序列。激光辅助显微切割技术（LCM）已经成功应用于分离积累在特异组织中的序列（尤其是稀有的转录本）。当用于分离的材料数量有限时，可用放大的方法来增加产物的数量。RNA 一旦转化为 cDNA，就可以作为 454 测序的模板获取表达序列标签（EST），用在随后的 SNP 分析和检测中（第 11 章，方案六）。

## 植物组织切片的激光辅助显微切割技术

激光辅助显微切割镜检已经发展成从不同类型的组织中获取样品的一项技术（Emmert-Buck et al. 1996；Simone et al. 1998），并已成功应用到植物样品中（Asano et al. 2002；Kerk et al. 2003；Nakazono et al. 2003）。很多不同的激光辅助显微切割平台都是商业性的，建议使用者结合自己的样品来评价这些平台。这章介绍的适用于玉米茎尖分生组织（SAM）的方案是根据 Asano 等（2002）和 KerK 等（2003）的方案改编并应用于 PALM 系统（P. A. L. M. Microlaser Technologies，Carl Zeiss，Bernried，Germany）。激光辅助显微切割和激光强制弹射（LMPC）技术使用脉冲紫外（UV-A）激光束切割来自组织截面的细胞并分配分离的组织到收集架中。

# RNA 放大

收集到标准显微切割中的 RNA 的数量常常不足以用于全基因组的表达分析，但它的数量却能够通过线性放大来增加。一个在 Eberwine 等（1992）方案基础上进行改进的程序用 oligo（dT）-T7 的联合引物通过两个循环依次经过逆转录和 RNA 转录来优先地选择多聚腺苷酸化的 RNA（如 mRNA）放大。这种放大是可重复的（Nakazono et al. 2003），并且典型的扩增数量在 50 000～500 000 倍（假设开始的材料为 10 ng 的 RNA，其中有 1%的 poly（A），放大后一般能产生 5～10 μg RNA 产物）。

## 方案一　玉米组织的制备和从目标细胞中提取 RNA

这个方案描述了丙酮固定的玉米幼苗组织石蜡切片的制备。一旦切片制备好，PALM 微光束系统，包括它的激光辅助显微切割和激光强制弹射技术就被用来切割感兴趣的细胞并将分离的组织弹到收集架中。然后这些组织就可以用来提取 RNA 了。

# 材料

注意：带有＜!＞标识的试剂要小心操作，具体见附录。

### 试剂

丙酮（100%，Fisher Scientific）＜!＞，冰冻和室温（RT）
焦碳酸二乙酯（DEPC；Sigma）＜!＞
玉米幼苗，14 d 大小
蜡（paraplast ＋，56℃，Oxford Labware）
RNA 提取缓冲液（举例来说，XB 来自 PicoPure RNA 分离试剂盒，Arcturus）
二甲苯（Fisher Scientific）＜!＞

### 仪器

解剖显微镜
细笔刷
Fisher 含探针的载玻片
梯度金属暖盘，也可用石蜡包埋中心
冰
金属称量盘
微量离心管（500 μL，带盖）
恒温箱（60℃）
PALM 微光束系统（PALM 微束 15VZ，P. A. L. M. 微激光技术，Carl Zeiss，http://www.palm-microlaser.com/dasat/index.php?cid=1001068conid=0&sid=dasat）

纸巾
石蜡储存器（2.5L，电子显微镜科研用）
玻璃培养皿
塑料袋
单面刀片
转轮切片机 Leica RM2135，Leica 显微系统或相同的设备
旋转器
手术刀
闪烁瓶（20 mL，Fisher Scientific）
载玻片加热盘（Fisher Scientific）
面巾纸
Tissue Tek 包埋环（电子显微镜科研用）
Tissue Tek 金属底模具（22×22×6，电子显微镜科研用）
真空器
称量勺

## 方法

### 固定

用冰冻的 100％丙酮作固定剂。固定应该在通风橱中进行。

（1）用单面刀片将幼苗在胚芽鞘节点处切断，立即放入装有冰丙醇的玻璃培养皿中。随即进行第二次切割用来收集茎尖生长点组织，切割点大约在胚芽鞘节点上端 1cm 的地方。

（2）整理浸没在固定剂中的含有 SAM 的组织到最终大小：0.3 cm×0.3 cm×0.2 cm。将它们放到有 25 mL 100％冰乙酸的闪烁瓶中，将闪烁瓶放在冰上。以这种方式准备 8～10 株幼苗并置于相同的管中，整个过程要迅速；每株幼苗的准备时间不应超过 10 min。

（3）将瓶放入 400 mmHg 的密闭空间中 10～15 min，对样品（在冰上）进行真空过滤。缓慢地使其平衡到大气压。用新鲜的 100％冰丙酮代替固定剂并将小瓶放在旋转器上在 4℃条件下旋转 1 h。

（4）将组织全部转移到带盖玻璃平皿中，将玻璃平皿放在解剖显微镜下，除去多余的叶子并用手术刀除去多余的节组织。最终的组织块大小应该接近于 0.3 cm × 0.2 cm×0.1 cm。重复步骤（3）两次（共 3 次过滤）。4℃条件下在旋转器上储存样品过夜。

### 脱水/二甲苯浸润

（5）恢复样品到室温，用室温下新鲜的 100％丙酮代替固定剂，室温下旋转 1 h。

（6）用丙酮∶二甲苯为 1∶1 的混合液取代固定剂并在室温下旋转 1.5 h。目标组

织非常稠密或隐藏在其他组织层中，用丙酮：二甲苯为 3：1、1：1、1：3 的混合液浸润。

(7) 用纯二甲苯洗 3 次，每洗一次保温 1 h。

## 蜡（paraplast）浸润

(8) 加入少量（通常为 10～15 小块，为小瓶体积的 1/10～1/5）的蜡碎屑到每个小瓶中，室温下在旋转器上培养过夜。

(9) 将小瓶都放到 60℃的恒温箱中以熔化剩下的蜡碎屑。当所有碎屑都熔化后，缓缓地倒置小瓶以混匀二甲苯和蜡。

(10) 60℃保温小瓶 1.5 h，加入更多蜡碎屑（加到小瓶一半体积），60℃保温 1.5 h。

(11) 用石蜡储存器中熔化的纯石蜡替换一半的二甲苯/蜡混合物，60℃保温过夜。

(12) 用纯蜡代替蜡混合物并在 60℃下保温。每天至少三次（每 4 h 一次）更换蜡。最后一次更换后 60℃下保温过夜。

(13) 第二天早上更换蜡并在 60℃下保温 4 h。

## 包埋

(14) 组装 Tissue Tek 金属底模具和 Tissue Tek 包埋环。准备一个由一端热而另一端是室温的梯度金属暖盘组成的“自制”包埋设备，或者使用石蜡包埋中心（如果有的话）。从 2.5L 的储存器中将蜡分配到组装好的带有包埋环的模具中。

(15) 将闪烁瓶中的组织和蜡倒入金属称量盘，称量盘已经放在暖盘的热端。用一称量勺将组织舀入组装好的带有包埋环的底模具中。摆正组织的位置以在旋转切片机上进行切片。将切片冷却至室温后放在冰上以便从模子中取出。在 4℃下储存在塑料袋中。

## 切片

(16) 水平切蜡块的顶端和底端，斜切侧端使其成为窄的梯形。将小块（10 μm 厚）放在旋转切片机上。

(17) 将带状片段放在 Fisher 含探针载玻片上，使其浮在 DEPC 水溶液中，40℃下在载玻片加热盘中加热 5 min，或直至切片伸展（不应超过 20 min）。

(18) 持带状片段的一端用细笔刷蘸载玻片到纸巾上以除去水，用面巾纸去掉残留的水。在除水的过程中不能使载玻片的降温。迅速地将载玻片放回载玻片加热盘直至完全干燥，或过夜。在 PALM 操作前一直在 4℃下储存干燥的载玻片。

## PALM 操作和 RNA 提取

(19) 加热切片至室温并通过 2 次二甲苯洗脱蜡（每次 10 min）。将一个载玻片从二甲苯中取出，允许其自然风干。把它置于显微镜台上标记要收集的组织所在区域。剩下的载玻片留在二甲苯中待用。

(20) 将约 50 μL RNA 提取缓冲液（如来自 PicoPure RNA 分离试剂盒的 XB，Arcturus）填入 500 μL 微量离心管的管帽中并加入 PALM 用来收集（见疑难解答）。

(21) 在收集细胞之前，将适合能量的激光聚焦于含有目标细胞的切片。

(22) 用 PALM 软件标记含有目标细胞的区域。

(23) 通过卖者的手册上“接近和切割 AutoLPC”法收集目标细胞。用聚焦的激光切出目标细胞的轮廓，分散的激光将已有轮廓的目标细胞弹射到切片上装有 RNA 提取缓冲液的管帽中。

每一种组织类型都要通过实验设置适合 PALM 操作系数。

(24) 收集足够数量的组织后，根据选择的方案提取 RNA（如来自 PicoPure RNA 分离试剂盒的 XB，Arcturus）。

6～10 个玉米茎尖分生组织（SAMs，含 15 000～18 000 个细胞）能提供约 10 ng RNA。

## 疑难解答

问题 [步骤（20）]：RNA 提取缓冲液干透了。

解答：在干燥的空气条件下一些 RNA 提取缓冲液会很快干透。晶体会在收集管帽的外面形成并会最终接触到载玻片。这可能会使剩下的液体缓冲液流到载玻片而损坏样品。为了避免这种情况的产生，可以用约 40 mL 的矿物油来代替 RNA 提取缓冲液（M. Scanlon，pers. Comm.）。收集组织到矿物油中，然后加入 RNA 提取缓冲液，充分混合。剩余部分的 RNA 提取缓冲液不会受矿物油的影响。

## 方案二 基于 T7 的玉米茎尖端分生组织 RNA 的扩增

这个方案描述了茎尖端分生组织（SAM）细胞 RNA 的扩增。它运用 oligo（dT）-T7 联合引物，通过两个循环依次经过逆转录和 RNA 转录优先地选择多聚腺苷酸化的 RNA。这个方案能够在 2～4 d 完成。

## 材料

注意：带有<!>标识的试剂要小心操作，具体见附录。

### 试剂

β-巯基乙醇（Acros）

β-烟酰胺腺嘌呤二核苷酸水合物（β-$NAD^+$；260（μmol/L），min. 98%来自酵母菌，Sigma）

氯仿（0.75%乙醇作为防腐剂，工业级纯，Fisher Scientific）<!>

焦碳酸二乙酯（DEPC）（Sigma）<!>

dNTP Set（10 mmol/L，Intermountain Scientific）

*E. coli* DNA 连接酶（10 U/μL，New England Biolabs）

*E. coli* DNA 聚合酶 I (10 U/μL, New England Biolabs)

带有 10×DNA 聚合酶缓冲液

乙醇（纯，Aaper 酒精）

MEGAscript T7 Kit (Ambion)

40 个反应；包括 rNTP 溶液，10× 反应缓冲液，T7 RNA 聚合酶混合酶，不含 RNA 酶的 DNA 酶 I

苯酚（饱和的，Fisher Scientific）<!>

步骤（10）：pH 6.6，BP1750I-400，步骤 18：pH 4.3，BP1751I-400

QIAquick PCR Purification Kit (Qiagen)

250 个柱；包括 PB、PE、EB 缓冲液

随机六聚体引物（1 μg/μL，Roche Diagnostics）

核糖核酸酶 H（RNase H；2 U/μL，Invitrogen）

提取自激光显微切割（LM）样品的 RNA

RNaseOUT 重组核酶抑制剂（40 U/μL，Invitrogen）

RNeasy Mini Kit (Qiagen)：50 个柱

包括 1.5 mL 和 2.0 mL 收集管和不含 RNA 酶的试剂和缓冲液

乙酸钠（100 mmol/L，pH 5.2，certified ACS，Fisher Scientific）<!>

SuperScript Ⅱ 逆转录酶（200 U/μL，Invitrogen）

带有 5×第一链合成缓冲液和 0.1 mol/L DTT <!>

T4 DNA 聚合酶（3 U/μL，New England Biolabs）

T4 基因 32 蛋白质（5 μg/μL，USB Corporation）

T7-oligo (dT) 引物（0.5 μg/μL，HPLC 纯化，DNA 整合技术公司（Integrated DNA Technolegies））

(5′-TC…TTTTT-3′)

### 仪器

浓缩器/蒸发器（Labconco CentriVap DNA 系统，Fisher Scientific）

加热台或水浴锅预设到 16℃、37℃、42℃、65℃、70℃、95℃

冰

微量离心管（无核酸酶）

漩涡搅拌器

## 方法

所有离心步骤都在室温下台式离心机中进行。

### 第一个循环的 RNA 扩增

（1）将下列复合物在一个不含核酸酶的离心管中混合：

| | |
|---|---|
| 0.5 μg/μL T7-oligo（dT）引物 | 1 μL |
| LM 样品中提取的总 DNA | 5～100 ng |
| 加水（经 DEPC 处理）至 | 11 μL |

（2）将样品在 65℃下保温 10 min 然后在冰上冷却 5～10 min。

（3）离心收集样品并使其平衡到 42℃保持 5 min。

（4）加入 8 μL 下列混合物到每个小管中：

| | |
|---|---|
| 10 mmol/L dNTP | 1 μL |
| 5×第一链合成缓冲液 | 4 μL |
| 0.1 mol/L DTT | 2 μL |
| 40 U/μL RNaseOUT | 0.5 μL |
| 5 μg/μL T4 基因 32 蛋白质 | 0.5 μL |

（5）缓慢混匀，并向每个小管中加 1 μL 反转录酶 Superscript Ⅱ（200 U/μL）。

（6）42℃保温 1 h

此时，如果需要，可以在－20℃下储存样品。

（7）每 20 μL 反应物中加入 130 μL 下列混合物

| | |
|---|---|
| 10×*E. Coli* DNA 聚合酶 I 缓冲液 | 15 μL |
| 10 mmol/L dNTPs 混合物 | 3 μL |
| 260 μmol/L β-$NAD^+$ | 15 μL |
| 10 U/μL *E. coli* DNA 聚合酶 I | 4 μL |
| 2 U/μL RNase H | 1 μL |
| 10 U/μL *E. coli* DNA 连接酶 | 1 μL |
| $H_2O$ | 91 μL |

（8）缓慢混匀，并在 16℃保温 2 h。

（9）加入 2 μL T4 DNA 聚合酶（3 U/μL）并在 16℃保温 10 min。

（10）用等量的 1∶1 苯酚（pH 6.6）/氯仿 提取双链 DNA。

（11）用等体积的氯仿提取并将水相转移到一支新的离心管中。

（12）用 Qiagen QIAquick PCR 纯化柱按下列步骤纯化 DNA

i. 向每个管中加入 35 μL100 mmol/L 乙酸钠溶液（pH5.2）

ii. 向每个管中分别加入 500 μLPB 缓冲液混合搅匀。

iii. 按照产商的指导逐次进行直到洗脱为止。

iv. 向每个柱中加入 15 μL $H_2O$，使其停留 1 min，以最大转速离心 1 min，重复此操作。

（13）在 50℃浓缩/蒸发仪中将样品浓缩至 8 μL。

（14）用 MEGA script T7 Kit 制备反应物

i. 将 rNTP 溶液解冻，旋转混合，离心并收集后置于冰上。

ii. 解冻 10×反应缓冲液，混合至沉淀物溶解，并保存于室温（非冰上）。

（15）将 20 μL 反应体积按照下列顺序加入

| | |
|---|---|
| cDNA | 8 μL |

| | |
|---|---|
| rNTP 混合物（ATP、CTP、GTP、UTP 各 2 μL） | 8 μL |
| 10×反应缓冲液 | 2 μL |
| T7 RNA 聚合酶混合物 | 2 μL |

(16) 将反应混合物于 37℃保温 5 h。

(17) 加入 1 μL 无 RNA 酶的 DNA 酶（2U/μL）并于 37℃下恒温保温 15 min。

(18) 向样品中加入 80 μL 无核酸酶的水，并用等体积 100 μL 1∶1 苯酚（pH 4.3）/氯仿溶液萃取。

(19) 用等体积的氯仿萃取并将水相转移至一个新的微离心管。

(20) 在 Rneasy mini column 中将样品浓缩。

i. 加入 350 μL RLT 缓冲液（及 3.5 μL β-巯基乙醇），并完全搅拌混匀。

ii. 加入 250 μL 无水乙醇并将其彻底吹吸混匀，无须离心。

iii. 将整个样品（700 μL）加入 2 mL 的 RNeasy minicolumn 收集管中。

iv. 以 10 000 r/min 速离心 15 s。

v. 将上述 RNeasy column 转移到另一个 2 mL 收集管中。

vi. 从 RNeasy column 吸取 500 μL RPE 缓冲液。

vii. 将其在 10 000 r/min 下离心 15 s，弃上清。

viii. 另向 RNeasy column 中加入 500 μL RPE 缓冲液，于 10 000 r/min 转速下离心 2 min，干燥 RNeasy 硅胶膜，弃上清。

ix. 将 RNeasy column 转移到另一个 1.5 mL 收集管中，并吸取 15 μL $H_2O$ 加入 RNeasy column。

x. 使 column 保持 1 min，然后并于 10 000 r/min 速离心 1 min。

xi. 另加 15 μL $H_2O$ 至 RNeasy column，保持 1 min，10 000 r/min 离心 1 min。

(21) 在浓缩/蒸发仪中将扩增的 RNA（aRNA）样品在 50℃条件下浓缩至 10 μL，若这一阶段需要对 RNA 进行定量，则将样品浓缩至 11 μL，并取 1 μL 做定量分析。

## 第二轮 RNA 扩增

(22) 混合 1 μL 随机六聚体引物（1 μg/μL）和 10 μL aRNA 以收集第一链反应产物，并于 70℃保存 10 min，后于冰上冷却 5 min。

(23) 离心收集样品，将离心管置于室温下平衡 10 min。

(24) 分别将 8 μL 下列混合物加至各管：

| | |
|---|---|
| 10 mmol/L dNTP 混合物 | 1 μL |
| 5×第一链合成缓冲液 | 4 μL |
| 0.1 mol/L DTT | 2 μL |
| 40U/μLRNaseOut | 0.5 μL |
| 5 μg/μL T4 基因 32 蛋白质 | 0.5 μL |

(25) 轻缓和混匀，加入 1 μL 反转录酶 Superscript Ⅱ（200U/μL），37℃保温放置 1 h。

(26) 加入 1 μL RNase H（2U/μL），于 37℃保温放置 30 min。

(27) 95℃加热 2 min，在冰上冷却样品 5 min。

(28) 加入 1 μL 0.5 μg/μL 的 T7-oligo (dT) 引物于 70℃保温放置 5 min。

(29) 在 42℃保温放置 10 min，然后置于冰上 5 min。

(30) 向每个管中加入 128 μL 下列混合物：

| | |
|---|---|
| 10×*E. coli* DNA 聚合酶Ⅰ缓冲液 | 15 μL |
| 10 mmol/L dNTP 混合物 | 3 μL |
| 260 μmol/L β-$NAD^+$ | 15 μL |
| 10U/μL *E. coli* DNA 聚合酶Ⅰ | 4 μL |
| 2U/μL RNase H | 1 μL |
| $H_2O$ | 90 μL |

(31) 重复第一轮 RNA 扩增中的步骤 8～20，viii。

(32) 洗提 aRNA，转移 RNeasy column 至另一 1.5 mL 收集管，吸取 30 μL $H_2O$ 至 RNeasy column，保持 1 min，10 000 r/min 离心 1 min。

(33) 另吸取 30 μL $H_2O$ 至 RNeasy column，保持 1 min，并于 10 000 r/min 离心 1 min。

(34) 用 1 μL 等分试样做 RNA 定量分析

经过纯化离心的 RNA 可用于 cDNA 合成，作为 454 转录组测序的模板（第 11 章，实验 6）。

## 致谢

我们感谢 Schnable 实验室人员和校友 Mikio Nakazono、David Skibbe 以及 Marianne Smith 对本章节中所列出的实验方案的贡献。本实验项目受到国家科学基金会植物基因组计划（DBI-0321595）的竞争性资助项目和哈奇法以及 State of lowa funds to P. S. S. 的支持。

## 参考文献

Asano T., Masumura T., Kusano H., Kikuchi S., Kurita A., Shimada H., and Kadowaki K. 2002. Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: Toward comprehensive analysis of the genes expressed in the rice phloem. *Plant J.* **32:** 401–408.

Eberwine J., Yeh H., Miyashiro K., Cao Y., Nair S., Finnell R., Zettel M., and Coleman P. 1992. Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci.* **89:** 3010–3014.

Emmert-Buck M.R., Bonner R.F., Smith P.D., Chuaqui R.F., Zhuang Z., Goldstein S.R., Weiss R.A., and Liotta L.A. 1996. Laser capture microdissection. *Science* **274:** 998–1001.

Kerk N.M., Ceserani T., Tausta S.L., Sussex I.M., and Nelson T.M. 2003. Laser capture microdissection of cells from plant tissues. *Plant Physiol.* **132:** 27–35.

Nakazono M., Qiu F., Borsuk L.A., and Schnable P.S. 2003. Laser-capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: Identification of genes expressed differentially in epidermal cells or vascular tissues of maize. *Plant Cell* **15:** 583–596.

Simone N.L., Bonner R.F., Gillespie J.W., Emmert-Buck M.R., and Liotta L.A. 1998. Laser-capture microdissection: Opening the microscopic frontier to molecular analysis. *Trends Genet.* **14:** 272–276.

# 9 哺乳动物 DNA 制备

Amrik Sahota,[1] Andrew I. Brooks,[1,2] and Jay A. Tischfield[1]

[1]*Rutgers University Cell and DNA Repository, Department of Genetics, Piscataway, New Jersey 08854-8082;*
[2]*Bionomics Research and Technology Center, Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635*



## 简介

从大量的个体获得高质量 DNA 是进行遗传多样性研究的先决条件。这一需要大大地刺激了 DNA 提取方法的技术进步。20 年前，DNA 大量的人工制备要花费超过 3 d

的时间才能完成。而现在运用自动化的提取设备——不需要使用有毒试剂就会较先前取得更大产量，大剂量的制备在 3 h 之内就可以完成。我们的实验室提出一种中大样品量（血液原液 3～30 mL 或 $1\times10^7\sim2\times10^8$ 培养细胞）DNA 提取的无机盐析方法（Madisen et al. 1987；Lahiri et al. 1992）。这种方法避免了苯酚对实验室人员以及环境的危害，且所有的试剂都是在室内制备的。我们运用这种方法已超过 15 年，但是，从便利性、高获得率和可信度的角度出发，近 5 年我们逐渐转向了商业试剂。比较研究显示，运用实验室人工提取方法获得的 DNA 与运用商业试剂的品质相当。商业试剂的缺点是实验人员对于试剂的化学成分以及他们所用的工具知之甚少。我们现在运用实验室程序作为备份并且将其展示于此，提供给那些认为购买商业试剂或者自动化设备过于昂贵的实验室。

## 组织的选择

进行一项研究最终所需的 DNA 数量是不可预测的。因此，我们倾向于首先从淋巴母细胞系（LCL）直接析取 DNA 然后从血液中提取。人类的 LCL 是利用 EB 病毒建立起来的（Neitze 1986；Caputo et al. 1991）。这些来源的 DNA 已经在多种基因型中心运用多种分析平台成功经过化验（Alarcon et al. 2002；Holmans et al. 2004；Prasad et al. 2005；Suarez et al. 2006）。随着近来两项科学技术的发展，即全基因组扩增（WGA）和阵列分析，似乎表明大量的基因组 DNA 并非总是必要的（Dean et al. 2002；Langmore 2002；Fathallah-Shaykh 2005；Dunphy 2006；Li et al. 2006；Quachenbush 2006；Reis-Filho et al. 2001；Tomlins et al. 2006）。

小量 DNA 的提取有很多种选择（如小于 0.5 mL 的血液、干燥血斑、唾液、口腔细胞、漱口水）（McEwen and Reilly 1994；Lum and Le Marchand 1998；Feigelson et al. 2001；Garcia-Closas et al. 2001；Heath et al. 2001；Steinberg et al. 2002）。可通过非侵害方法收集获得的样品（如唾液），于血样相比更加可行（Steinberg et al. 2002；Rylander-Rudqvist et al. 2006）。口腔擦洗物和漱口水已被用于许多流行病学研究，但对于 DNA 提取量，尤其是口腔擦洗物 DNA 提取量明显低于唾液样品。正如本章节所述，保存在 Oregene 溶液（DNA genotek，渥太华，安大略湖，加拿大）中的唾液较漱口水样品更易运输。许多由口腔擦洗物和漱口水提取 DNA 的实验方法已经被收集出版（Lum and Le Marchand 1998；Feigelson et al. 2001；Garcia-Closas et al. 2001；Heath et al. 2001；Steinberg et al. 2002）。

## 抗凝血剂的选择

EDTA 是 DNA 提取中血样收集的抗凝血剂，它抑制脱氧核糖核苷酶的活性且不引起体积变化。ACD 管可用于血液收集，更重要的是它将血液充满至标记线从而避免引起渗透压的变化，后者可引起细胞降解。肝磷脂应当去除，它在纯化阶段可结合到 DNA，并抑制用于 PCR 反应的 Taq 酶的活性。若不考虑抗凝血剂，真空采血管需倒转若干次以混合血液。血液可以在环境温度下运送，但是如果收集和提取多于 3 d 的话，那么 DNA 会有所降解，产量低于新鲜血液。

## 提取小结

所有DNA提取的步骤都包括细胞裂解、除去蛋白质和DNA的再溶解。根据样品种类的不同，也可能包括除去RNA的环节。例如，当DNA从血液中分离出来的时候，无细胞核的红细胞将首先被裂解，以从白细胞中分离出来。而白细胞随后被一种阴离子降解剂裂解，后者将细胞组分溶解。这种作用被一种能够抑制脱氧核糖核苷酶活性的DNA防腐剂所阻断。细胞质和细胞核蛋白可被盐沉淀转移。基因组DNA被乙醇沉淀并溶解于TE缓冲液。部分DNA被稀释到特定的浓度，提取的DNA和稀释样品分成小份储存在－70℃。这可以避免反复的冷冻和大量样品的溶解，从而防止沉淀形成。为了防止DNA的酸性水解，在此不用纯水来储藏。水中的任何杂质也都可能引起DNA的降解。

## DNA质和量

为了进行DNA精确定量，在0.1～1.0，应当运用吸光度测定法进行测定。若超出此范围则应当调整样品稀释参数。同样，由于吸光度测定法和$A_{260/280}$比率都是与pH相关（Willinger et al. 1997），因而该测定法也应相应运用于碱性缓冲液（10 mmol/L Tris-HCl，1 mL EDTA，pH 8.0）。较低的pH会引起DNA的降解，而且会造成较低的$A_{260/280}$比率和降低对蛋白质污染物的敏感度。由于水的pH浮动范围过大，因此水不可用于此项检验。另外，相关计算参数（$IA_{260}=50\ \mu g/mL$ DAN）也是基于pH的，当用水进行稀释时参数并非严格正确。

按照常规DNA浓度，由UV光谱仪测定。这种方法廉价、迅速、持久且为半自动化，但它所测量的是UV吸收物质的总量，这包括从细胞系分离出的DNA夹带的RNA（从血液中提取的DNA中只有一点或者没有RNA残余）。因此有些情况下，UV测量法可能低估DNA的含量。当DNA的量有限时，UV吸光度可以通过NanoDrop ND-1000分光光度计（Nano-Drop科技，威尔明顿，特拉华州）测定。这种仪器仅需要1 μL的样品量，可不经稀释而测定。

自动化的程序，如那些基于荧光性的程序也可以用于DNA浓度的测定。基于荧光性的方法较UV测量法具有更高的灵敏性和特异性，如PicoGreen法，但是检测方法的选择主要取决于特定样品的浓度范围。因此，PicoGreen并不适于测定“储存”DNA浓度，而UV测量法则更适合测定“工作”DNA的浓度。换句话讲，定量PCR（qPCR）可用于测定样品浓度极低的DNA浓度，如法医研究或者源于固定组织提取的DNA测定（见方案十二）。

1％琼脂糖凝胶电泳显示运用本章中所描述方法分离出的DNA通常大于23 kb，因此适合于Southern印迹和其他分子生物学研究，同样适用于PCR（Lahiri et al. 1992）。DNA产量取决于材料大小、类型、年龄和质量。由小细胞所构成的样品，如转化的淋巴细胞，将会比由大细胞组成的同样大小的样品具有更高的细胞密度，因此有更高的DNA含量。

### 蛋白质、RNA 和其他杂质

质量良好的 DNA 的 $A_{260/280}$ 比率应该在 1.8～2.0。若一个样品的 $A_{260/280}$ 比率低于 1.6（预示蛋白质杂质的污染），则应对其重新进行沉淀以改善纯度（Manchester 1995，1996；G-lasel 1997）。若 $A_{260/280}$ 比率高于 2.0 则标志着 RNA 杂质的存在，但对于血液样品来说这通常并不成问题。当分析合并的 DNA 池样品（如直接测定等位基因频率）时需要确定各个样品具有相等数量的 DNA 时，DNA 酶 A 的处理就尤其重要。DNA 中的 RNA 杂质会在凝胶电泳时的染色剂前端呈现出弥散条带。我们已经发现苯酚-氯仿提取很少需要进一步的 DNA 纯化，当我们开始使用 Gentra 自动化萃取程序后就更不需要了。PCR 扩增显示由唾液提取的 DNA 混有支原体 DNA（A. Sahota et al.　unpubl.）和不定量的细菌 DNA，这些都会干扰基因型分析检测，尤其是多元平台的检测。这些杂质在血液或者细胞系 DNA 中都未曾观察到。

### 限制酶的剪切

作为进一步的质量检测，我们用两种不同的限制性内切核酸酶来降解各 DNA 样品并且利用已知被这些酶彻底降解的 DNA 样品（细胞系 K562）作为对照来判断样品 DNA 的降解程度。我们分别使用 6 碱基（*Eco*RⅠ）识别序列和 4 碱基（*Hae*Ⅲ）识别序列限制性内切核酸酶。降解产物片段在琼脂糖凝胶上分布在相对分子质量大小不同的区域。已经观察到运用上述一或两种酶的部分剪切所得到的超出预期片断大小范围的更高相对分子质量的片段。运用 *Eco*RⅠ的完全酶切表明所用的 DNA 不含盐类杂质，因为这种酶十分容易受到盐类浓度变化的影响。所有未被彻底剪切的 DNA 样品都要用高盐和异丙醇重新沉淀纯化，再重复上述限制酶剪切过程。

本章展示了一系列各种来源，包括细胞系、组织、血液和唾液 DNA 的提纯。本章也涵盖了全基因组扩增和法医研究中 DNA 提取的方法。其中所提到的商品和公司名称并非意味着作者对其的宣传；其他公司的产品也同样有效。最终的 DNA 样品质量检测有多种方法，取决于各实验室。运用这里所述的方法制备的 DNA 已为全世界很多实验室的基因型研究作出贡献，未出现任何问题（Alarcon et al.　2002；Holmans et al.　2004；Prasad et al.　2005；Suarez et al.　2006）。

## 致谢

我们感谢下列同事在科学和技术上的贡献：Douglas A. Fugman，David A. Toke，Qi Wang，Laura Wilde，David Keller，and Hiep Tran。Rutgers 大学细胞与 DNA 研究组（RUCDR）的工作由国家研究院（NHI）和诸多私立组织承认签约支持。RUCDR 作为国家细胞研究组致力于酒精中毒遗传学合作项目（NIAAA）、精神失调合作研究中心（NIMH）、遗传学研究中心（NIDA）、遗传研究所（NIDDK）和细胞与 DNA 研究所从而服务于诸多私立机构。

## 方案一　细胞沉淀物 DNA 制备

Amrik Sahota,[1] Andrew I. Brooks,[1,2] and Jay A. Tischfield[1]

*[1]Rutgers University Cell and DNA Repository, Department of Genetics, Piscataway, New Jersey 08854-8082; [2]Bionomics Research and Technology Center, Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635*

本实验描述了运用内部或者商业（PUREGENE）试剂从细胞沉淀物中提取大量DNA 的方法。这些细胞沉淀物是从已经用 EB 病毒建立的人类淋巴母细胞系（LCL）中制备的。

## 材料

注意：带有<!>标识的试剂要小心操作，具体见附录。

### 试剂

琼脂糖凝胶（1%）
氯仿<!>
乙醇（70%）
溴化二氨乙苯啡啶<!>
人体 LCLs
异丙醇<!>
NaCl（0.85%，6 mol/L，饱和溶液）
蛋白酶 K（41.7 mg/mL 水中原液）<!>
PUREGENE 试剂（Gentra 系统）（商业试剂方案）
　细胞裂解液（D-50K2）
　DNA 溶解液（D-50K4）
　蛋白质沉淀液（D-50K3）
　蛋白酶 K（D-50K5）
　RNase A 溶液（D-50K6）
限制性内切核酸酶：*Eco*RⅠ和 *Hae*Ⅲ
RNase A（无脱氧核糖核苷酶，10 mg/mL）<!>
含有 15%乳牛血清（FCS）的 RPMI1640 培养基
SDS（20%）<!>
TE 缓冲液（1×）（10 mmol/L Tris-HCl，1 mmol/L EDTA [pH 8.0]）
TKM #2 缓冲液（10 mmol/L Tris-HCl，10 mmol/L KCl，10 mmol/L $MgCl_2 \cdot 6H_2O$，2 mmol/L EDTA，0.4 mol/L NaCl [pH 8.0]）

## 仪器

带有GH-3.8转子的Allegra GS-6离心机（Beckman），或类似的台式离心机
生物废物缸
离心管若干（50 mL圆锥形）
低温瓶若干（8 mL；1 mL和2 mL具条形码）
鱼钩（由一种巴斯德玻璃移液管制成）
烧瓶若干（T-175）
事先设置为37℃和56℃的恒温箱
轨道混合器
摇床
分光光度计
Vi-CELL细胞发育分析仪（Beckman Coulter）
涡旋混合器

# 方法1

## 应用实验自制试剂提取DNA

(1) 在T-175烧瓶中加入含有15%乳牛血清（FCS）的RPMI1640作为LCL培养基，扩增细胞至大于$1.5\times10^8$。为了保证质量，随机选择15%～20%的培养基用Vi-CELL细胞发育分析仪来检测细胞的数量和细胞生活力。

(2) 将各培养液分别收集到50 mL圆锥形离心管中并以3600 r/min离心5 min，收集细胞。用0.85%的NaCl溶液清洗细胞团两次以清除培养基以及血清蛋白。

(3) 对于一个包含约$2\times10^8$细胞的细胞团来说，加入12.8 mL的TKM #2缓冲液和400 μL的20%SDS，再加入50 μL RNase A。将样品于37℃恒温放置1 h。

根据细胞团的大小调整试剂量。

(4) 加入100 μL蛋白酶K并于56℃保温放置3 h。此后，将样品置于室温下过夜或者度过周末。

(5) 向样品中加入4.8 mL的6 mol/L NaCl饱和溶液（最终浓度为1.6 mol/L）。将管子涡旋混合15 s后，在室温下放置5 min变性沉淀蛋白质。以3000 r/min离心10 min后，将上层水相转移到另一个50 mL圆锥形离心管中，而不要接触到沉淀的蛋白质。

(6) 向收集的上清液中加入等量室温下的异丙醇，颠倒离心管若干次以使DNA沉淀。用鱼钩（由一种玻璃的巴斯德吸液管制成）将DNA链从异丙醇中拉出并用70%乙醇漂洗两次。将尖端附有DNA的鱼钩转移到8 mL的低温瓶并向其中加入4 mL 1×TE缓冲液。然后向瓶中加入一滴氯仿置于摇床上直到DNA完全溶解（需过几天）。

(7) 用分光光度计测量溶液在260 nm及280 nm处的吸光度值以估算DNA的浓度。可以将分光光度计设定为自动测定DNA溶液浓度、产量和$A_{260/280}$比率，将数据传

输到 DNA 数据库中。

(8) 按如下方式检验 DNA 质量

i. 用限制性内切核酸酶 *Eco*RⅠ和 *Hae*Ⅲ（在不同的反应中）剪切 1 μL DNA。

ii. 将上述产物进行 1%琼脂糖凝胶电泳。

iii. 用溴化乙啶对凝胶染色。

[应当没有或只出现少量未切割 DNA，同时要用未剪切的 DNA 等分样品进行凝胶电泳，以验证其是否具有较大的分子质量（>23 kb）。]

(9) 将制备的 DNA 平均分置于 2 个或多个具条形码的低温瓶中，储存于－70℃。为减少储存 DNA 的冻融次数到最低，用 1×TE 缓冲液将其稀释至 100 ng/μL（或者其他特定浓度），并各取 300 μL（30 μg DNA）置于 6 个 1 mL 具条形码的低温瓶中，于－70℃储存直至进一步分析。

[T-175 培养基物（10 μg/$10^6$ 细胞）的 DNA 产量为 1500～2000 μg，其 $A_{260/280}$ 比率与血液制备的 DNA 样品相似，且容易被 *Eco*RⅠ和 *Hae*Ⅲ剪切。]

## 方法 2

### 用商业试剂提取 DNA

(1) 准备 0.5×$10^8$～1.5×$10^8$ 个已转化的淋巴母细胞。按照方法 1 中步骤（2）描述的方法，将细胞在盐溶液中漂洗，并在剩余的盐溶液中重悬细胞团。加入 15 mL 含有 RNase（75 μL）的细胞裂解液，涡流震荡 30 s 以裂解细胞。

(2) 将样品置于 37℃保温 15 min，冷却至室温后快速离心以收集所有液体至管底部。

(3) 加入 5 mL 蛋白质沉淀液，涡流震荡 15 s，3600 r/min 离心 15 min 沉淀蛋白质。将含有 DNA 的表面液体倒入新的洁净离心管中，加入 15 mL 异丙醇，将离心管颠倒 20 次以沉淀 DNA。

(4) 将异丙醇倒入废液缸，倒置离心管排除水分 60 s。向其中加入 15 mL 70%乙醇并轻柔混合以漂洗 DNA。将样品以 4200 r/min 离心 15 min 以收集 DNA。倒掉乙醇，并倒置离心管 60 s，空气烘干。

(5) 加入 4 mL DNA 溶解液，并于 56℃放置于轨道式混合器中 1 h，使 DNA 充分溶解。最后将 DNA 溶液转移到 8 mL 低温瓶中旋转若干天。

(6) 按照方法 1 中的步骤（7）、(8) 来测定 DNA 的含量和质量，将 DNA 按照步骤（9）储藏。

## 方案二 从固定组织制备 DNA：提取与全基因组扩增

Chad Brueck, Clyde Brown, Steve Michalik, Deborah Vassar-Nieto, Ernie Mueller, and Gary Davis

*Sigma-Aldrich, St. Louis, Missouri 63103*

虽然样品中存在 DNA 降解水平难以恢复的特殊问题，但是基因组 DNA 仍然可以

从固定的组织中提取。全基因组扩增技术提供了来自新鲜、冷冻、储存的，或在经过化学处理保持基因完整性材料的 DNA 高效扩增和永久保存的方法。这种方法的显著优势是能够扩增降解或者数量非常有限的 DNA。

由福尔马林固定、石蜡仓埋的组织（FFPE）几十年来在医疗检验过程中获得了丰富的材料，成为这种核酸的一个潜在来源。这些现成可用的临床材料具有相应临床和病理数据，从中可恢复可用基因组 DNA。关于福尔马林的固定对 DNA 造成破坏的解释有多种假说。最常见的解释为 DNA 经过羟甲基化作用或者碱基不可逆的交联而被进一步降解。另外，已知甲醛可以直接通过半肽键将蛋白质的某些基团，如硫醇、酚或终端氨基基团交联于 DNA。另一个假说更离谱，它假设这种固定的方法聚合蛋白质结合在 DNA 上，导致 DNA 聚合物难于与模板结合，因此使得这种序列用多种扩增方法都不能恢复。

GenomePlex WGA 系统特别适合这些具有挑战性的降解样品（方案 9 中 GenomePlex WGA 策略的描述）。Tissue WGA 试剂盒特别用来于 1 mL 组织中提取的降解了的基因组 DNA 的扩增。然而，像 0.1 mL 这样少量的组织也用这项产品成功地扩增。在这个方案中，在 EFPE 组织恒温保存 1 h，进行组织降解、释放基因组 DNA 过程中加入了蛋白酶 K。在这期间，DNA 通过化学和酶两种方法被部分修复了。然后，一份溶解产物可直接用于 WGA 反应。此试剂盒避免了应用有机溶剂去除石蜡及产生平均 5～10 μg WGA 产物。提纯过后，可以用其他基因组或者染色体组 DNA 样品同样的方法分析 WGA 产品（图 9-1 定量 PCR 分析）。

本实验方案用于组织中基因组成分的提取和序列扩增。

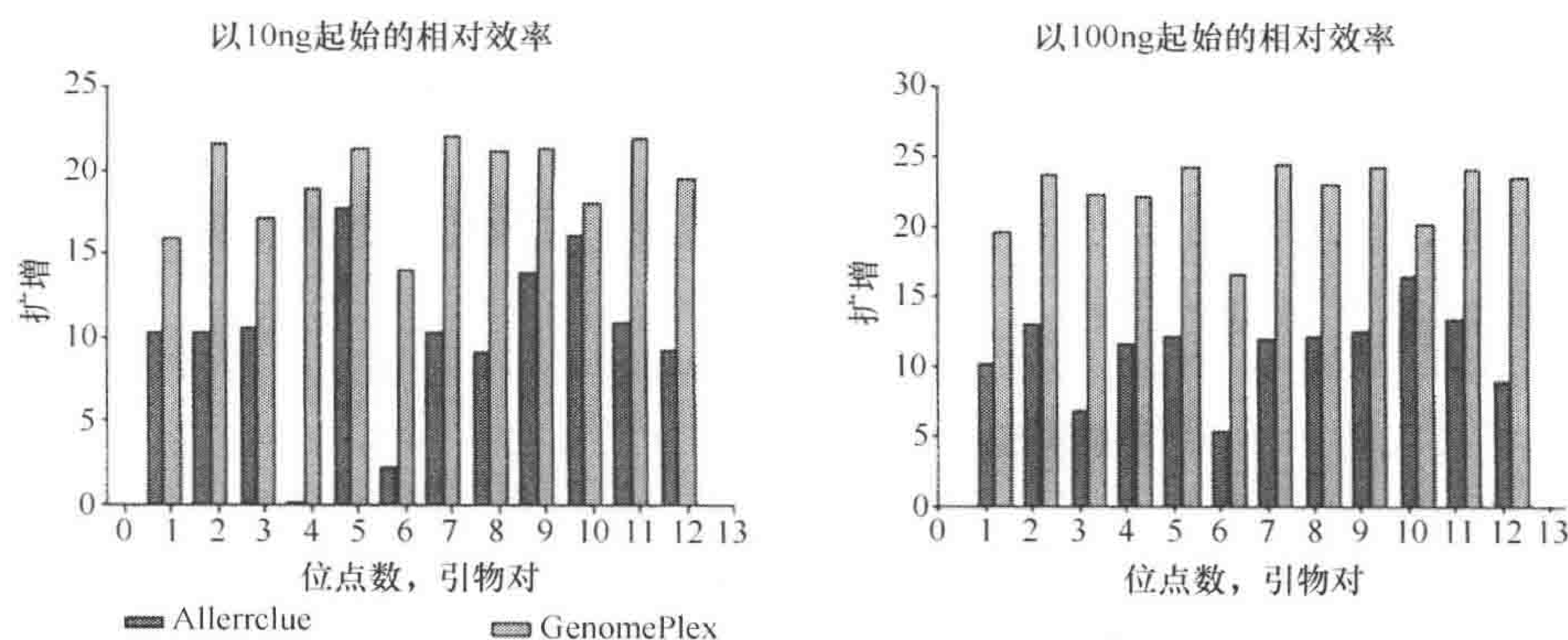



图 9-1　FFPE 样品的结果。本实验测定在扩增受损（FFPE）DNA 过程中的 PEP/DOP 和 MDA 两种相对扩增效率

# 材料

## 试剂

新鲜、冷冻或者 FFPE 组织（块状或者片体）

GenomePlex 组织全基因组扩增试剂盒（Sigma），包括

扩增母液混合液（10×）

细胞溶解酶 Y 裂解液

文库制备缓冲液

文库制备酶

文库稳定液

去核酸酶的 $H_2O$

蛋白酶 K 溶液

WGA 聚合酶

### 仪器

用来称量毫克单位组织的天平

移液器和枪头（专用的）

安全刀片

分光光度计

PCR 仪（PE 9700 或者类似产品）

薄壁 PCR 管（0.2 mL 或者 0.5 mL），或者 96 孔 PCR 平板

## 方法

### 裂解和破碎

（1）称量 1 mg 组织样品置于 PCR 管。当作用于 FFPE 组织时，用刀片将多余的石蜡除去。

[微量到 0.1 mg 的组织使用该试剂盒也能成功。但最适合的最小微量组织量应通过实验确定通常从 1 mg 开始。]

（2）向样品中加入 24 μL 细胞溶解酶 Y 裂解液和 6 μL 蛋白酶 K 溶液。

（3）将反应物于 60℃保温放置 1 h，然后于 99℃放置 4 min。

[该组织经过蛋白酶 K 作用可能并未降解，GenomePlex 组织全基因组扩增试剂盒不必将组织完全降解。非完全降解的组织就可以提取到用于 WGA 足够的 DNA。]

（4）立即置于冰上冷却样品，然后快速、短暂离心以收集产物。

（5）将 9 μL 去核酸酶的 $H_2O$ 与 1 μL 步骤（4）得到的组织溶解产物混合置于一个洁净的 PCR 管。

（6）加入 2 μL 1× 文库制备溶液。

（7）加 1 μL 文库稳定液。

（8）彻底混匀并将其于 95℃ PCR 仪中保温放置 2 min。

（9）立即于冰上冷却，离心收集后放回冰上。

（10）加入 1 μL 文库制备酶，充分搅拌，快速离心。

（11）将样品放置于 PCR 仪中按下列程序保温

i. 16℃20 min

ii. 24℃20 min

iii. 37℃20 min

iv. 75℃5 min

v. 保持在 4℃

(12) 将样品从 PCR 仪中取出并快速离心。样品立即扩增或者于－20℃储存 3 d 扩增。

(13) 向步骤 (12) 的 14 μL 反应体系中加入：

7.5 μL 扩增母液混合物 (10×)

48.5 μL 无核酸酶的 $H_2O$

5 μL WGA DNA 聚合酶

(14) 充分混匀，快速离心，并开始 PCR 反应。下列 PE 9700 或者 PCR 仪参数已经被优化设定为

| 循环数 | 变性 | 退火/延伸 |
| --- | --- | --- |
| 第 1 次循环 | 95℃2 min | |
| 20 次循环 | 94℃15 s | 65℃4 min |

(15) 循环完成后，于 4℃维持反应或于－20℃储藏直到纯化和序列分析。推荐用 Sigma 的 PCR Cleanup 试剂盒进行 WGA DNA 纯化。WGA DNA 与同等条件下储存的基因组 DNA 的稳定性没有差别。事先进行样品纯化对于 DNA 定量十分必要，因为这样可以去除能干扰后续操作的未结合引物。

## 疑难解答

问题：扩增后产量很低。

解决方法：

(1) WGA 反应可能被提取物中的杂质抑制。用 $H_2O$ 以 1∶10 或更大比例稀释组织溶液中的抑制物后继续制备文库。否则，就是用了过多的组织材料。为防止这种情况出现，我们尽量用较少的组织。为检测抑制，将对照 DNA 和 (或) 10 ng 纯化的基因组 DNA 模板加入蛋白酶 K 降解液中。

(2) 称量少于 0.1～1 mg 的组织。事先用刀片除去多余石蜡再对组织进行称重。这样就可保证在随后的蛋白酶 K 酶解液中有足够量的组织。

(3) FFPE 组织中 DNA 模板质量可能会较低。如果固定过程时间过长或者 FFPE 样品储藏不当，福尔马林固定会造成 DNA 不可逆的损伤。尽量获得固定得当的 FFPE 组织样品。

(4) 提取物可能不足。使样品于 60℃保温放置样品超过 1 h，可保温样品 2 h 以过夜，然后在 99℃保温 4 min。

(5) 反应后纯化可能不恰当。用一种方法只保留单链或双链 DNA。推荐使用 Sig-

ma PCR Cleanup 试剂盒。

(6) 蛋白酶 K 酶解后的组织可能未完全消化。GenomePlex 组织全基因组扩增试剂盒不需要组织被完全降解。未完全降解组织，也可能提取适量的 DNA 用于 WGA 扩增。

问题：WGA 扩增的目标基因的 qPCR 反应表现出强烈的偏好。

解决方案：

(1) 对照设置不恰当。一旦对照 DNA 被剪切，基因组 DNA 就只能与 WGA DNA 比对。解决方法是运用几种经过降解 DNA 方案获得的混合样品，或者经过水解的 DNA 作对照 (Thorstenson et al. 1998)。

(2) DNA 样品量太少或过分降解。见“低产量”解决方案。

问题：阴性（无模板）对照组生成产物。

解决方案：一种或多种试剂可能被外源 DNA 污染，解决的办法是替换这批反应试剂（尽管这个问题也许并不影响结果，但一个干净的无模板对照结果只能通过替换反应试剂来获得）。

## 方案三　从大鼠尾或耳分离 DNA

Edwin Cuppen

*Hubrecht Laboratory, Utrecht, The Netherlands*

本实验描述从大鼠尾或耳分离 DNA 的快速方法。此实验方案最简便的版本可用于 96 孔（深孔）平板。DNA 数量适合任何基于 PCR 的基因型加工。

## 材料

注意：参见附录指导对标记＜!＞的试剂要小心操作

### 试剂

乙醇 (70%)

异丙醇＜!＞

苯酚/氯仿 (1∶1) ＜!＞ [可选，见步骤 (5)]

蛋白酶 K 裂解缓冲液

100 mmol/L Tris-HCl (pH 8.5)、200 mmol/L NaCl、0.2% SDS＜!＞、5 mmol/L EDTA

加入新鲜 10 mg/mL 蛋白酶 K 储存液 ＜!＞（储存于−20℃）体积的 1/100

TE (10 mmol/L Tris，pH 7.5、0.1 mmol/L EDTA) [可选，见步骤 (13)]

组织样品 (0.25cm 尾或者耳组织片)

### 仪器

调整到 55℃和 80℃的恒温箱或水浴

试管若干或 96 孔（深孔）平板

涡旋混合器

## 方法

（1）收集后立刻将组织样品转移到装有 400 μL 新鲜蛋白质 K 裂解缓冲液的管中。

（2）闭管或者将 96 孔板密封。涡旋混合。

（3）55℃或者保温放置至少 4 h，保温过夜。时常涡旋混匀。

（4）于 80℃加热 15 min 以使蛋白酶 K 变性失活。冷却平板离心，收集冷凝物。

下面步骤（5）～（7）是可选择的（与平板方案不同），要获得更纯的 DNA，这些步骤是必要的。

（5）加入 400 μL 苯酚/氯仿（1∶1）液。

（6）涡旋旋转混合 1 min。最高转速离心（试管：14 000*g*；平板：6000*g*）5 min。

（7）将上清液转移到一个新试管或平板中。

重复上述步骤直到界面上看不到蛋白质/SDS 白色沉淀物。

（8）加入 300 μL 异丙醇并颠倒混匀 10 min。

（9）最高转速离心（试管：14 000*g* 10 min，平板：6000*g*）40 min。

（10）通过倒置试管或平板弃去上清液。用 300 μL 70％乙醇漂洗（不需涡旋混合）。

（11）最高转速离心（试管：14 000*g* 5 min，平板：6000*g*）15 min。

（12）通过倒置试管或平板弃去上清液。空气干燥 DNA 平板。

（13）将 DNA 沉淀物溶解于 500 μL 的 MilliQ $H_2O$ 或者 TE 中。

### 方案四　从口腔细胞制备 DNA

Irena B. King

*Molecular Diagnostics Program, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109*

流行病学研究所需的 DNA 通常来自采集的血液样本。然而，这种方法被认为对参与者有侵害性并且对于大规模的研究来讲十分昂贵（King et al. 2002）。对采集的其他样本包括毛发、指甲、唾液和口腔细胞的研究逐渐发展。提取口腔细胞可用吹洗法、擦洗或者用一种轻柔的细胞刷获得。本实验描述了利用细胞刷收集口腔细胞提取 DNA 的过程。在利用适合多种生物原始材料的亲和柱基础上，为增加相关组织的 DNA 产量对最初的步骤进行了优化。这个方法并不高产，获得高产的潜在危害是较低的 DNA 增加相关，导致分析数据质量波动。因此，这个实验方案需要额外的质量保证和对照检验。实验之前，先了解 QIAamp DNA 微提取试剂盒的工作手册。

## 细胞刷收集样品

一般来说，给定样品所能分析出的基因型数目受到原始材料的数量和质量的制约。血液样品能提供足量的高品质的 DNA，但是由细胞刷，尤其是通过邮寄获得样品的

DNA，会有更大程度的降解并含有不定量的细菌 DNA。尽管细菌 DNA 并不干扰 PCR 成功进行，但用于特殊用途的口腔细胞 DNA 需要验证确认（Nicklas and Buel 2003）。细胞刷获取 DNA 的另一个弊端是其有限的产量。尽管用于增加 DNA 数量的全基因组扩增实验已经成功地进行了至少 10 年，但是关于实验保真度的问题一直都是关注的焦点（Lee et al. 2006）。

细胞刷采样需要先漱口然后用细胞刷与颊内侧缓和摩擦。清晨在刷牙和吃东西之前取样，通常能获得略多于其他时间取样的 DNA。为增加 DNA 总量又不增加对参与者的损害，选择每次收集 3 刷口腔细胞。这个数量对于参与者和每个细胞刷上等量 DNA 的产量来说已被普遍接受。尽管还不清楚完成口腔黏膜修复需要多长时间，但最好是在多次样品收集之间留足至少一周时间。从一个细胞刷上面获得的总 DNA 量平均为 3 μg，若参试人员积极主动且训练有素，其产量可略有上升（King et al. 2002）。如表 9-1 所示，在最近提取 VITAL cohort 研究中（见致谢），1002 个参与者细胞刷提取的 DNA，总产量为 0.1～30 μg/细胞刷。

**表 9-1　邮购的 1002 个口腔细胞刷样品 DNA 含量总结**

| | 浓度/(ng/μL) | DNA 总产量/μg |
|---|---|---|
| 最小量 | 0.5 | 0.1 |
| 1% | 1 | 0.2 |
| 10% | 5 | 0.8 |
| 25% | 10 | 1.5 |
| 中位数 | 18 | 2.6 |
| 平均值 | 22 | 3.3 |
| 75% | 29 | 4.3 |
| 90% | 43 | 6.4 |
| 99% | 88 | 13.2 |
| 最大量 | 201 | 30.2 |

# 材料

## 试剂

无水乙醇

磷酸盐缓冲溶液（PBS）

QIAamp DNA 微量提取试剂盒溶液（Qiagen）（按照 QIAamp DNA 微提试剂盒说明书配制）

AE 缓冲液（作为补充）

AL 缓冲液（于暗处室温储存）

AW1 缓冲液（室温储存）

AW2 缓冲液（室温储存）

蛋白酶 K 溶液

大马哈鱼精细胞DNA标本（0～250 ng/μL溶解于去降DNase $H_2O$中）（Sigma）

待分析样本

对于大规模的研究，含有自我收集的口腔细胞的细胞刷常常需要邮局及时处理和大批量递送。2～3 d收集的细胞刷邮递样品仍然可用。运用条码标签进行邮寄以便快速浏览返回的样本。在邮寄之前，将样品冷冻，尤其当扫描和邮递不在同一天进行时更需如此。扫描过后，将细胞刷样本按25个一包于－80℃储藏在一个特制的箱子中(2″×2″×6.25″）直至分析。如果按照上述方法操作，储藏至少5年，其DNA的数量和质量都不会发生明显的改变。

## 仪器

连接数据库的条形码读码器（如果用条形码标签）

储存细胞刷样品的箱子（2″×2″×6.25″）

聚丙烯圆锥管（15 mL，去降DNase）

细胞刷（Medical Packaging公司）

标签（每个样品5个，白色高级激光印刷）（Island Scientific）

微型离心管（1.5 mL，去降DNase）

微移液器和枪头（100 μL和1000 μL）

平板阅读器（如SpectraMax 250，分子装置）

Quartz微型板（96孔）供UV读取

QIAamp DNA微提试剂盒（250）（Qiagen）

如果没有QIAamp试剂盒，类似的产品也能获得相当的产量，如Gentra的Puregene

15 mL离心管和微型离心管架

剪刀

面巾纸

涡旋混合器

56℃水浴锅

# 方法

（1）按如下方式从细胞刷提取DNA。

i. 将含有样品的细胞刷放置于15 mL的聚丙烯锥形离心管中，剪去其塑料把手，留下距离其顶端大约1″的长度。向管中加入400 μL PBS，旋转，于室温下保存15 min。

ii. 加入20 μL蛋白酶K溶液和400 μL AL缓冲液。立即涡旋混匀30 s。

iii. 将锥形管置于56℃水浴中保存30 min。室温下将锥形管以1800*g*离心1 min 。每次离心之前都需用面巾纸将离心管擦拭干净，以防止交叉污染。

iv. 加入400 μL 100%乙醇并涡旋混匀。室温下再次以1800*g*离心1 min。

（2）按如下方式将细胞刷毛上多余的DNA除去。

i. 将锥形管中的细胞刷移出并将其小心放入一个1000 μL移液枪头中，将移液枪头放入另一个15 mL的锥形管，将黏膜细胞混合物室温保存在第一个锥形管中。

ii. 将移液枪头中含有黏膜细胞的第二个锥形管以 1800$g$ 离心 5 min，使多余 DNA 通过移液枪头开口从细胞刷转移到锥形管中。小心移动并弃去含有细胞刷的移液枪头。

(3) 将含有黏膜细胞的 700 μL 混合物小心地由第一个锥形管转移到一个旋转收集柱中（试剂盒提供），并在室温下 6800$g$ 离心收集柱 1 min。将旋转收集柱从微型离心机取出，放在一个 2 mL 收集管中。

(4) 将第一个、第二个锥形管中剩余的成分混合，转移到准备好的旋转收集柱。再次以 6800$g$ 微离心 1 min。

(5) 小心将 500 μL AW1 缓冲液加到收集柱，6800$g$ 离心 1 min。

(6) 弃去漂洗液并将旋转收集柱（含 DNA）放置于一个新的 2 mL 收集管（试剂盒提供）中。小心地将 500 μL AW1 缓冲液加入柱内，6800$g$ 离心 1 min。

(7) 将旋转收集柱转移到一个新的收集管中并加入 500 μL AW2 缓冲液，注意切勿弄湿管口边缘。以 20 800$g$ 对含有旋转收集柱的收集管离心 3 min。再加入 500 μL AW2 缓冲液，重复上述离心 1 min。

(8) 弃去含有漂洗液的收集管并将旋转收集柱置于 1.5 mL 离心管中。吸取 150 μL AE 缓冲液（洗提缓冲液）至柱上。室温下放置 10 min，6800$g$ 离心 1 min。

(9) 将洗提缓冲液（含 DNA）从离心管中吸出，转移到一个旋转柱（此步骤可优化 DNA 产量），再次以 6800$g$ 离心 1 min。

(10) 弃去旋转柱，将洗提后的 DNA 存于贴有标签并盖紧的微型离心管中，−80℃长期保存，−20℃短期保存。

(11) 检测 DNA 浓度确定其产量。具体方法取决于应用的需要。以下方法可供选择：

i. 荧光染色 pico 绿（Invitrogen）。该方法花费最大，但高度精确灵敏，样品需要量最少。

ii. 在琼脂糖凝胶上参照标准梯度条带比较谱带强度。此方法最不灵敏，但提供了 DNA 大小（DNA 质量）的信息。

iii. 建立 6 点标定曲线，用鲑鱼精细胞标准 DNA（0～250 ng/μL）以分光光度法检测 DNA 量。测 10 μL 样品（用无 DNase 水 1∶20 稀释）260 nm 和 280 nm 吸光度。$OD_{260}/OD_{280}$ 率 $R$（DNA 质量）应当满足 1.6～2.0。

通过吸光度进行的 DNA 定量适用于多数 PCR 及其后续工作。吸收单位应降到 0.1～1.0，较低的 DNA 浓度的检测需要使用其他的技术，如 pico 绿染色。

## 致谢

Irate King 感谢 Emily White 博士提供的 1002 细胞刷样品的 DNA 产量分布资料，她是 VITAL 研究的主要负责人。VITAL（VITamin and Lifestyle）是对华盛顿西部 75 000名 50～74 岁男女进行的营养保健品和癌症风险人群的研究。

## 方案五 全血基因组 DNA 制备：中量、小量提取

Amrik Sahota,[1] Andrew I. Brooks,[1,2] Jay A. Tischfield,[1] and Irena B. King[3]

*[1]Rutgers University Cell and DNA Repository, Department of Genetics, Piscataway, New Jersey 08854-8082; [2]Bionomics Research and Technology Center, Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635; [3]Molecular Diagnostics Program, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109*

本方案介绍了使用 QIAamp DNA 血样小提试剂盒（200 μL 样品）或 QIAamp DNA 血液样中提试剂盒（0.5 mL 样品）从全血样品中提取基因组 DNA 的方法。方案基于对亲和的使用，适用于多种不同生物起始材料，为获得 DNA 最大产量初始步骤根据 DNA 的不同组织来源进行了优化。本过程不适用于大量提取。该方案用于大量提取会造成 DNA 低复原率及由此导致的分析数据质量波动是本方案进行大量提取的潜在障碍。这样的过程要求附加的质量保证及对照检验。按此方案得到的 DNA 产量取决于血液样品中有核细胞的数量以及冻存样品的质量。小量提取的理论产量为每 200 μL 血样 4～12 μg，每 200 μL 白细胞层 25～50 μg（Qiagen 数据）。健康供血者 0.5 mL 血样的产量是 8～35 μg。提取 DNA 之前，应该先熟悉 QIAamp DNA 血液小提或中提试剂盒使用手册。DNA 库必须强调高度保密，按监督部门制定的规章制度，实行详细的列表归档保存，所有血液样品应按照危险品须知和管理手册 Hazard Awareness Management Manual（HAMM）中的详细规定处理。

## 材料

### 试剂

琼脂糖凝胶

乙醇（纯）

磷酸盐缓冲液（PBS，无菌）

QIAamp DNA 血液中提试剂盒（Qiagen）（按照 QIAamp DNA 血液中提试剂盒使用手册准备）

AE 缓冲液（用作补充）

DNA 样品需规格化的还需额外添加（见步骤 10iii）

AL 缓冲液（暗处室温保存）

AW1 缓冲液（室温保存）

AW2 缓冲液（室温保存）

QIAGEN 蛋白酶储存溶液（4℃，可保存 2～3 月）

QIAamp DNA 血液小提试剂盒（用于小量提取）

AE 缓冲液（用作补充）

下列三种缓冲液室温可保存 1 年

AL 缓冲液（使用前充分摇匀）

AW1 缓冲液（加入 25 mL 96%～100%乙醇备用）

AW2 缓冲液（加入 30 mL 96%～100%乙醇备用）

蛋白酶

用移液器将 1.2 mL 蛋白酶溶液（含 0.04%叠氮化钠的去核酸酶水）加入冻干 Qiagen 蛋白酶中。将蛋白酶分为 20 μL 等份存于－20℃，以延长试剂保存期。

待分析样品：

小量提取：使用 200 μL 血样或白细胞层样品。

中量提取：准备新鲜抗凝血样等分后的－80℃冻存样品。EDTA 和肝磷脂可能会影响 PCR 反应（Garcia et al. 2002）。每批提取 16～20 份样品。每日可提两批或以上，依实验安排而定。

### 器材

连接数据库的条形码阅读器（如需使用条形码标签）

适用 15 mL 锥形试管的离心机（常温）

聚丙烯锥形试管（15 mL，每份样品 2 支）

冻存盒（9×9），81 瓶

琼脂糖凝胶电泳装置

56℃电热板或水浴装置（用于小量提取）

微量离心管（1.5 mL 除脱氧核糖核酸酶，每份样品 4 支）

微型移液器和枪头（100 μL、500 μL、1 mL）

酶标仪（如 SpectraMax 250、Molecular Device）

石英微孔板（96 孔）供紫外测量

QIAamp DNA 血液中提试剂盒（100）（中量提取；Qiagen）

QIAamp DNA 血液小提试剂盒（小量提取；Qiagen）

15 mL 和微型离心试管架

涡旋振荡器

37℃和 70℃水浴装置

# 方法 1

## 小量 DNA 提取

(A. Sahota, A.I. Brooks, and J.A. Tischfield)

（1）56℃在缓冲液 AL 中溶解所有沉淀。向装有 20 μL Qiagen 蛋白酶（或蛋白酶 K）的微型离心管中加入 200 μL 样品（血样或白细胞层）。若样品体积不足 200 μL 再加入足够 PBS 使体积达到 200 μL。用移液器枪头抽吸搅匀。

推荐使用蛋白酶作为细胞裂解液，但蛋白酶 K 更适合含高浓度 EDTA 或 SDS 缓冲

液的组织样品。

(2) 样品中加入200 μL缓冲液AL，涡旋振荡15 s，水浴或电热板56℃保存10 min。将试管短暂离心使样品沉降到底部。加入200 μL 100%乙醇至样品中，涡旋振荡15 s。再经短暂离心去除残余液体。

(3) 向QIAamp旋转结合柱中加入混合液（620 μL），注意不要弄湿边缘。在微型离心机中以8000 r/min（6000g）离心1 min。将含有滤出液的收集管弃入内盛消毒剂（来苏水）的生物试剂废液缸再将结合柱转移到新的2 mL收集管中。加入500 μL缓冲液AW1，注意不要弄湿边缘。按上述的方法离心，弃去含滤出液的收集管，将结合柱转到新的2 mL收集管中。

(4) 加入500 μL缓冲液AW1，注意不要弄湿边缘。在微型离心机中以8000 r/min离心1 min，弃去含滤出液的收集管，将结合柱转移到新的2 mL收集管中。加入500 μL缓冲液AW2，注意不要弄湿边缘。在微型离心机中以14 000 r/min（20 000g）离心3 min，弃去含滤出液的收集管，将结合柱转移到新的2 mL收集管中。

(5) 在微型离心机中以14 000 r/min离心1 min除去缓冲液。将QIAamp旋转结合柱转移至一个干净的1.5 mL微型离心管中，弃去含残余滤出液的收集管。向柱中加入200 μL缓冲液AE洗提DNA。室温保存5 min，8000 r/min离心1 min。

(6) 检测DNA质量并通过PCR进行分析。使用时保持DNA在4℃以便使用，长期保存时要置于−70℃。

DNA的产量一般为每200 μL血样4～12 μg，每200 μL白细胞层25～50 μg（Qiagen数据）。Qiagen旋转结合柱对DNA和RNA都能纯化，这对血液样品不存在问题，因为RNA含量极低。如果需要除净RNA的DNA样品，可在加入缓冲液AL前加入20 μL RNase A溶液（20 mg/mL）。

# 方法2

## 中量DNA提取

(I.B. King)

(1) 37℃放置15 min解冻全血样品，冰浴直至实验开始。

抗凝全血储存样品必须迅速解冻再冷藏或冰浴。

(2) 用移液器将0.5 mL全血移至一支15 mL锥形聚丙烯离心管中，加入0.5 mL冷的无菌PBS。再向样品中加入100 μL Qiagen蛋白酶储存溶液和1.2 mL缓冲液AL。加盖振荡15 s使彻底混合，70℃水浴10 min。

(3) 样品停止水浴并加入1 mL乙醇。再次振荡15 s使混合。

(4) 小心将全部溶液转移至15 mL锥形试管的QIAamp中型柱。避免溢出或弄湿QIAamp中型柱的边缘，加盖并以1850g（约3000 r/min）离心3 min。

离心中每一步都要用纸巾擦拭试管防止交叉污染。

(5) 转移QIAamp中型柱，弃滤液，将QIAamp中型柱转移到原来的15 mL离心

管中。向 QIAamp 中型柱中小心加入 2 mL 缓冲液 AW1，避免弄湿边缘。加盖并以 4500g（约 5000 r/min）离心 1 min。

（6）向 QIAamp 中型柱中小心加入 2 mL 缓冲液 AW2，避免弄湿边缘。加盖并以 4500g（约 5000 r/min）离心 15 min。

（7）将 QIAamp 中型柱转移至一个干净的 15 mL 离心管中，弃去含滤出液的收集管。加入 100 μL 缓冲液 AE，放置至室温，直接加到 QIAamp 中型柱表面并加盖。室温放置 5 min。以 4500g（约 5000 r/min）离心 5 min。

（8）为获得最大 DNA 产量，用枪头吸取 100 μL 新鲜 AE 缓冲液，放置至室温，直接加到 QIAamp 中型柱表面。室温放置 5 min，加盖并以 4500g（约 5000 rpm）离心 5 min。弃去 QIAamp 中型柱，并小心盖好瓶盖。

（9）冷藏直至 DNA 呈粉状。

（10）90 份提取液（6 批）完成后，将提取的 DNA 分批进行微孔板质量检验。按以下步骤进行分析。

i. 微孔板 DNA 质量检验

ii. 每份样品取 50～100 ng 通过琼脂糖凝胶电泳进行 DNA 质量评估。对凝胶观察并作如下评分：0＝很好，1＝好，2＝不好。

iii. 为防止冻存/解冻速度重复，每一份样品分为 1～3 等份或统一标准化装瓶。按储存液终体积为 150 μL 计，分装后各等分按统一至 DNA 浓度为 30 ng/μL，每份 DNA 总量约 250 ng。计算缓冲液量加入样品，使 DNA 达到要求浓度并分装，以达上述标准。

当提取液 DNA 量低于上述正常量时，如果低于 30 ng/μL 应保持其原始浓度分装储存。

（11）按如下步骤准备冻存样品。

i. 在样品分装试管前，用预先印好的条码标签粘贴到微型离心管上。每一份样品单独编号。分批交叉检查贴标签的小瓶防止出错。对原初提取液采用特殊标签加以区分。

ii. 确保所有离心管瓶盖关紧。将每一等份竖直放置于 9×9 冰存盒中，从盒的左上角开始，顺序连续地在每孔中放一支小瓶。原初提取液与分装的样品要分别放置并在盒外清楚标记。标签命名惯例可为“学名，样品类型（全血 DNA），体积，小盒系列号”。按数据库资料核实小瓶编号顺序。

iii. －80℃冷藏以供长期保存。

DNA 提取、定量、标准化需要自始至终保有跟踪记录数据，并有严格的质量控制和复查。

## 方案六　血液 DNA 制备：大量提取

Amrik Sahota,[1] Andrew I. Brooks,[1,2] and Jay A. Tischfield[1]

*[1]Rutgers University Cell and DNA Repository, Department of Genetics, Piscataway, New Jersey 08854-8082; [2]Bionomics Research and Technology Center, Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635*

本方案描述了用于实验室或工业用途的血液 DNA 大量提取的方法，包括使用实验

室制备的或商用试剂（puregene），也包括一种从半凝固或凝固后的血样提取的 DNA 方法。

## 材料

注意：参见附录指导对标记＜！＞的试剂要小心操作。

### 试剂

琼脂糖凝胶（1%）
氯仿＜！＞
乙醇（70%）
溴化乙锭＜！＞
异丙醇＜！＞
氯化钠（0.85%）
氯化钠（6 mol/L 饱和溶液）
155 mmol/L 氯化铵＜！＞/170 mmol/L Tris（pH7.7）（预热至 37℃）
蛋白酶 K[4]（41.7 mg/mL 水溶液）＜！＞
PUREGENE 试剂（Gentra 系统）（用于商用试剂制备法及半凝固和凝固血样的制备方法）
　　细胞裂解液（D-50K2）
　　DNA 水溶液（D-50K4）
　　糖原溶液（20 mg/mL，R-5010）（用于半凝固和凝固血样制备方法）
　　蛋白质沉降液（D-50K3）
　　红血球裂解液（D-50K1）
　　RNase A（D-50K6）[可选，见方法 2 步骤（3）]
限制性内切酶：*Eco*RⅠ及 *Hae*Ⅲ
RNase A＜！＞（除净 DNase，10 mg/mL）[可选，见方法 1 步骤（5）]
SDS（20%）＜！＞
TE 缓冲液（1×）（10 mmol/L Tris-HCl，1 mmol/L EDTA [pH8.0]）
TKM＃2 裂解缓冲液（10 mmol/L Tris-HCl，10 mmol/L 氯化钾，10 mmol/L 六水氯化镁，2 mmol/L EDTA，0.4 mol/L 氯化钠，pH8.0）
EDTA 全血收集
　　实验室试剂制备方法：1～3 支试管，8 mL 试管为宜
　　商用试剂及半凝固和凝固血样的 DNA 提取方法：5～10 mL

### 器材

吸水纸（商用试剂制备方法）
Allegra GS-6 离心机配备 GH-3.8 转轴（Beckman）或同类型台式离心机

生物试剂废液缸（内盛消毒剂，如来苏水）
离心管（50 mL 锥形）
低温瓶（1 mL 及 2 mL，条码编号）
钓钩（用玻璃巴氏吸液管制成）
电热板预设 56℃（可选）
冰（用于半凝固和凝固血样）
恒温箱预设 56℃
定轨摇床
振动平台
分光光度计
移液器（塑料）
涡旋振荡器
37℃水浴装置

## 方法 1

### 实验室配置试剂的大量提取方法

(1) 从收集在 1～3 只试管中的 EDTA-全血中分离白细胞（白细胞层）。在 Allegra GS-6 离心机（配备 GH-3.8 转轴）或同类型台式离心机中以 3000 r/min 离心 10 min。

(2) 用一支塑料移液器将白细胞层转移到 50 mL 锥形离心管中。为在以后的生化实验中使用将上清液（包括血浆和血小板）保存在－70℃。将底层（含红细胞）倒入其中一支装血样的试管中，保存于－20℃或－70℃。

样品含足够的非成熟（有核）红细胞和白细胞用于小量 DNA 提取，可以用来解决任何样品鉴别问题。样品鉴别问题一旦解决储存样品可以舍弃。

(3) 向白细胞层中加入 40 mL 的 155 mmol/L 氯化铵/170 mmol/L Tris（pH7.7）（预热至 37℃），缓慢搅拌，37℃温水浴 7 min。

此过程裂解白细胞层中的红细胞并释放了血红蛋白。温度和时间关键。低的温度或更短的时间都无法让红细胞完全裂解。而水浴时间过长则会导致白细胞裂解发生，而导致 DNA 产量降低。

(4) 以 3000 r/min 离心 10 min 后将上清液倒入生物试剂废液缸。将白细胞置于 40 mL 0.85%NaCl 中，再离心，弃上清。

本步骤将剩余血红蛋白除尽。而且沉淀物中应呈纯白色，不会有红色素，不然的话产物 DNA 在提取后要进一步纯化。

(5) 将沉淀在 6.4 mL TKM＃2 裂解缓冲液重悬混匀，再加入 200 μL 20%SDS（最终浓度 0.6%）和 50 μL 蛋白酶 K（最终浓度 300 μg/mL）。在定轨摇床上用 56℃温水浴 1～2 h。之后样品可置于室温过夜或一周。RNase A 处理在全血 DNA 分离中通常并不必要。如果需要，在加入蛋白酶 K 前加入 25 μL 除净 DNAse 的 RNase A（10 mg/mL）达到 35 μg/mL 最终浓度，样品 37℃温水浴 30 min。

本方案试剂量通常用于约 20 mL 血中正常白细胞含量（每毫升个数 $4\times10^6$～$8\times10^6$），为近似 $10^8$ 个白细胞。根据沉淀物颗粒大小试剂用量也要相应进行增减。

（6）向样品中加入 2.4 mL 6 mol/L NaCl 溶液（最终浓度 1.6 mol/L）。涡旋振荡试管 15 s，于室温直立放置 5 min 使蛋白质变性沉降。3000 r/min 离心 10 min，将上清液转移至新的 50 mL 锥形离心试管中，不要搅动蛋白质沉淀物。

（7）向上清液中加入等量室温异丙醇，并将试管倒转数次使 DNA 沉降。用钓钩（玻璃巴氏吸液管制）将成股 DNA 从异丙醇中拉出，以 70%乙醇漂洗。将钓钩尖端连同 DNA 转移到一个 2 mL 条码编号的冻存瓶中，加入 1 mL 1×TE 缓冲液。然后向瓶中加入一滴氯仿再转移到振动平台直至 DNA 完全溶解（数日）。

（8）通过分光光度计测得 260 nm、280 nm 吸光值计算 DNA 浓度。设定程序令分光光度计自动计算 DNA 浓度、产量（基于原始血样量）以及 $A_{260/280}$ 比例，并上传数据至 DNA 数据库。

全血提取 DNA 的产量为 15～35 μg/mL（取决于白细胞含量），20 mL 的血样产量在 480 μg 左右，$A_{260/280}$ 比例在 1.9 左右。

（9）用下列方法检验 DNA 质量。

i. 用限制性内切核酸酶 *Eco*RⅠ和 *Hae*Ⅲ消化 1 μg DNA（独立的实验）。

ii. 在 1%琼脂糖凝胶中对消化产物进行电泳。

iii. 对凝胶进行溴化乙锭染色。

实验结果应显示没有或仅存少量未降解 DNA。凝胶上要有完全降解的等量 DNA 作为对照以检验是否有高分子 DNA 存在（>23 kb）。

（10）等分的储存 DNA 分装两个或多个条码编号冰存瓶中，存于－70℃。为使冰冻/解冻率降至最小，用 1×TE 缓冲液制备 100 ng/μL 稀释液分装 300 μL（300 μg DNA）至 6 个 1 mL 条形码编号冰存瓶，冻存于－70℃供给基因型分析中心、PCR 实验室分析等使用。

## 方法 2

### 使用商业试剂的 DNA 大量提取

有很多商业试剂盒可用于 DNA 大量提取。在进行广泛研究认证，我们采用 PUREGENE 实验程序进行血样 DNA 的大量提取。

（1）将 5～10 mLEDTA 全血转移至一支 50 mL 锥形离心管中。加入 RBC 裂解液（裂解液与血液的总量应达到 40 mL）。缓慢倒转试管数次使混合均匀。室温静置 7 min。

（2）Allegra GS-6 离心机或同类型台式离心机 3600 r/min 转速离心 5 min 收集白细胞。将上清液倒入废液缸中。重悬剩余液体中沉淀颗粒，加入 10 mL 细胞裂解液，用吸液管吹吸混合数次裂解细胞。样品可在细胞裂解液中保存至少 2 年。

（3）如果需要进行 RNase 处理，加入 50 μLRNase A（4 mg/mL）倒转离心管数次混匀，样品 37℃水浴 15 min。样品冷却至室温再加入 3.33 mL 蛋白质沉淀液。涡旋振荡 15 s 使沉淀液与细胞裂解产物混合。

(4) 样品以 3600 r/min 离心 5 min。沉淀的蛋白质将形成致密的暗色颗粒。将含 DNA 的上清液转入一支 50 mL 锥形离心管中并加入 10 mL 异丙醇。缓慢颠倒试管 20 次使 DNA 沉降。样品以 3600 r/min 离心 5 min 收集 DNA。将异丙醇上清倒入废液缸，将含 DNA 沉淀颗粒的试管倒置于一张干净的吸水纸上 60 s。

(5) 向 DNA 沉淀颗粒的试管中加入 10 mL70%乙醇，倒转试管数次漂洗颗粒。样品以 4200 r/min 离心 15 min 收集 DNA。将乙醇倒入废液缸，将试管倒置于干净吸水纸上 60 s，干燥数分钟。

(6) 加入 1 mL DNA 水溶解液将 DNA 颗粒再次溶解。在定轨摇床上 56℃温水浴 1 h。将 DNA 转移到 2 mL 冰存瓶中室温置于振动平台上直至 DNA 完全溶解（数日）。

(7) 测定 DNA 含量与质量后准备分装以便长期保存（方法 1 的 8～10 步）。

10 mL 血 DNA 理论产量 300 μg（在 150～500 μg），取决于白细胞数量。DNA 是大分子样可提取 DNA 约（>23 kb），可被多种限制性内切核酸酶降解，包括 *Eco*RⅠ、*Hae*Ⅲ、*Hin*DⅢ和 *Pst*Ⅰ。据 Gentra 报道 DNA 可在−80℃条件下保存至少 10 年。

# 方法 3

## 半凝固和凝固血样的 DNA 提取

半凝固血样包括：①采集超过 3 d 的血样；②储存于 4℃条件下超过 1 周的血样；③储存于−20～−70℃的血样；④采集时未适当混合或不正确采集到红顶瓶（凝集管）的血样。对于此类样品，基于使用 GentraSystems 实验方案的如下方法。

(1) 若半凝固血样冻结，在 37℃下迅速解冻，转移到 50 mL 锥形试管中。用 RBC 裂解液将体积扩充到 40 mL。均匀混合，室温放置 5 min，其间至少搅拌 1 次。样品以 3600 r/min 转速离心 5 min。吸去大部分上清液，试管中留下大约 4 mL 液体。涡旋振荡 30 s 使细胞再次悬于液体中，加入 10 mL 细胞裂解液，再次振荡。

(2) 37℃温水浴至少 2 h 或室温过夜，使样品呈现均匀状态。将样品在冰中冷却 5 min，加入 4.5 mL 蛋白沉淀液，振荡 20 s。3600 r/min 转速离心 10 min。将含 DNA 的上清液转移到干净的 50 mL 锥形离心管中并加入 13.5 mL 异丙醇和 133 μL（20 mg/mL）糖原溶液。倒转试管数次混合。

(3) 3600 r/min 转速离心 5 min 收集 DNA。弃去异丙醇，除去试管中水，用 70%乙醇漂洗沉淀颗粒。4200 r/min 转速离心 15 min，弃去乙醇，除去试管中水，空气干燥颗粒。用 500 μL DNA 水溶解液将 DNA 再次溶解。

(4) 对于凝固血样再增加以下步骤。

i. 加入 RBC 裂解液后，高速振荡试管再将其置于振荡器上，室温振荡 5 min 使凝结块尽量溶解。

ii. 3600 r/min 转速离心 5 min，弃上清，再使颗粒悬于 10 mLRBC 裂解液中。重复振荡、水浴步骤。

iii. 样品离心，弃上清，再使颗粒悬于约 400 μL 剩余液体中。

iv. 加入 10 mL 细胞裂解液和 50 μL 蛋白酶 K 溶液（20 mg/mL），涡旋振荡，56℃

水浴 2 h 或过夜，直至凝结块完全溶解。

v. 样品在冰中冷却 5 min，加入 3.3 mL 蛋白沉淀液，高速涡旋振荡。

vi. 3600 r/min 转速离心 10 min，取出样品置冰中冷却 5 min，将含 DNA 的上清液转移到一干净的 50 mL 锥形离心管中。

vii. 加入含糖原溶液 20 μL（20 mg/mL）的异丙醇 10 mL。

viii. 用 70%乙醇漂洗沉淀颗粒并用 500 μL DNA 水溶液将 DNA 再次溶解。

半凝固血样 DNA 提取量大约是新鲜血样的一半，主要取决冷冻解冻过程中 DNase 的作用和白细胞的裂解程度。凝固血样的 DNA 提取产量更低。

## 方案七　唾液 DNA 制备

Amrik Sahota,[1] Andrew I. Brooks,[1,2] and Jay A. Tischfield[1]

*[1]Rutgers University Cell and DNA Repository, Department of Genetics, Piscataway, New Jersey 08854-8082; [2]Bionomics Research and Technology Center, Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635*

唾液有时可作为血液替代物成为 DNA 提取源，其优势是可以不侵害人体而直接采集。本方案描述了从 2 mL 唾液样品中提取 DNA 的 Oragene 方法。它能为研究和诊断提供高质量 DNA。

## 材料

### 试剂

乙醇（95%～100%）
Oragene DNA 自提取试剂盒溶液（DNA Genotek）
TE 缓冲液

### 器材

Allegra GS-6 离心机或同类型台式离心机
离心试管（15 mL 锥形）
冰存瓶（2 mL，条码编号）
冰
恒温箱预设 50℃、56℃
微型离心管（1.5 mL）
Oragene DNA 自提取试剂盒（DNA Genotek）
分光光度计
定轨摇床
振动平台
涡旋振荡器

## 方法

（1）用试剂盒中的塑料容器收集 2 mL 唾液。将盛有唾液容器后装有 2 mL Oragene DNA 保藏液的小杯的封条拧开。封条打开防腐剂会进入唾液中。将容器颠倒数次混匀。

（2）唾液/Oragene 溶解产物可在普通环境温度转运、储存超过一年。如需长期保存，则需将溶解物分成小份，－20℃保藏于微型离心管中。

（3）溶解物 50℃保藏过夜以备 DNA 提取。

下列对 Oragene Purifier 的使用指导基于 Oragene 使用手册。

（4）如果需要全部样品的提取，将溶解物转移到一支 15 mL 锥形离心管中，加入 160 μL（1/25 溶解物量）Oragene Purifier 溶液。用涡旋振荡数秒混匀样品，冰浴 10 min使蛋白质和其他杂质沉降，用 Allegra GS-6 离心机或同类型台式离心机以 2500g 离心 10 min。将上清转移至一干净的 15 mL 锥形离心管中，注意不要扰动颗粒。弃去颗粒。

对少量溶解物则相应减少纯化溶液容量。

（5）估算上清液体积加入等量乙醇（95%～100%）。通过倒转离心管混合样品，试管在室温下放置 10 min 使 DNA 沉降。

由于浓度的不同，DNA 可能以纤维或理想的沉淀物形式出现。

（6）按步骤（4）离心样品并弃上清。如果需要，再次短暂离心去除剩余上清液。

（7）用 1 mL TE 缓冲液将 DNA 再次悬浮，将试管置于 56℃ 定轨摇床 1 h。将溶液转移到 1.5 mL 微型离心管中，室温置于振动平台上直至 DNA 完全溶解（至少1 d）。

（8）样品室温下 15 000g（13 000 r/min）离心 15 min 以除去所有微粒物质。将 DNA 溶液转移到一支 2 mL 条码编号的冰存瓶中。测定 DNA 的含量和质量。如需长期保藏，则将其均等地分装到微型离心管中，在－70℃冻存。

根据 DNA Genotek 提供的数据，DNA 再溶后浓度为 2～200 μg/mL，每 2 mL 唾液提取 DNA 的量在 110 μg 左右（15～300 μg）。$A_{260/280}$ 比例在 1.6 左右，但用 320 nm 吸光度校正后会增至 1.7 以上（取决于样品浊度）（见疑难解答）。

唾液/Oragene 混合物中提取的 DNA 也可用 PUREGENE 试剂、Qiagen 小柱法或其他方法。

## 疑难解答

问题［步骤（8）］：DNA 溶液出现浑浊。

解决：DNA 在 320 nm 下不吸收光，但若溶液出现浑浊，在此波长下进行测算校正浊度对应的吸光度。从 $A_{260}$ 和 $A_{280}$ 中减去 $A_{320}$ 从而得出正确的 $A_{260/280}$ 比率。这通常会使 $A_{260/280}$ 比率由 1.5 升至 1.7 以上。在对混浊样品计算 DNA 浓度过程中同样需要采取这种矫正。

## 方案八 用 PCR 对基因组 DNA 进行全基因组扩增

Chad Brueck, Clyde Brown, Steve Michalik, Deborah Vassar-Nieto, Ernie Mueller, and Gary Davis
*Sigma-Aldrich, St. Louis, Missouri 63103*

多数 DNA 分析需要微克量核酸作为起点。但是，对于来源有限的材料即使最好的 DNA 提取方法也难以产出足够的产物，而且即使起始分析步骤中 DNA 量足够，之后的检测也会很快耗尽少量的 DNA。这些问题使很多研究者致力于发展能够克服以上障碍的方法。

全基因组扩增（WGA）是上述问题的一种通用的解决办法。它提供了对来自新鲜、冷冻、保存或化学处理样品的基因组 DNA 的有效扩增和永久性储存的途径，并可以保证基因组 DNA 的遗传完整性。当前，两种全基因组扩增方法可商业化。第一种，多链移位扩增（MDA），使用一种高效的嗜热 DNA 多聚酶和随机引物建立由 20kb～70kb 产物（Dean et al. 2001）构成的完整基因组。第二种，使用引物延伸预扩增（Zhang et al. 1992）和兼并性寡核苷酸引物（DOP）（Telenius et al. 1992）PCR，制备由大约 440bp 短段组成的全基因组拷贝。方法详述如下（简要步骤见图 9-2）。

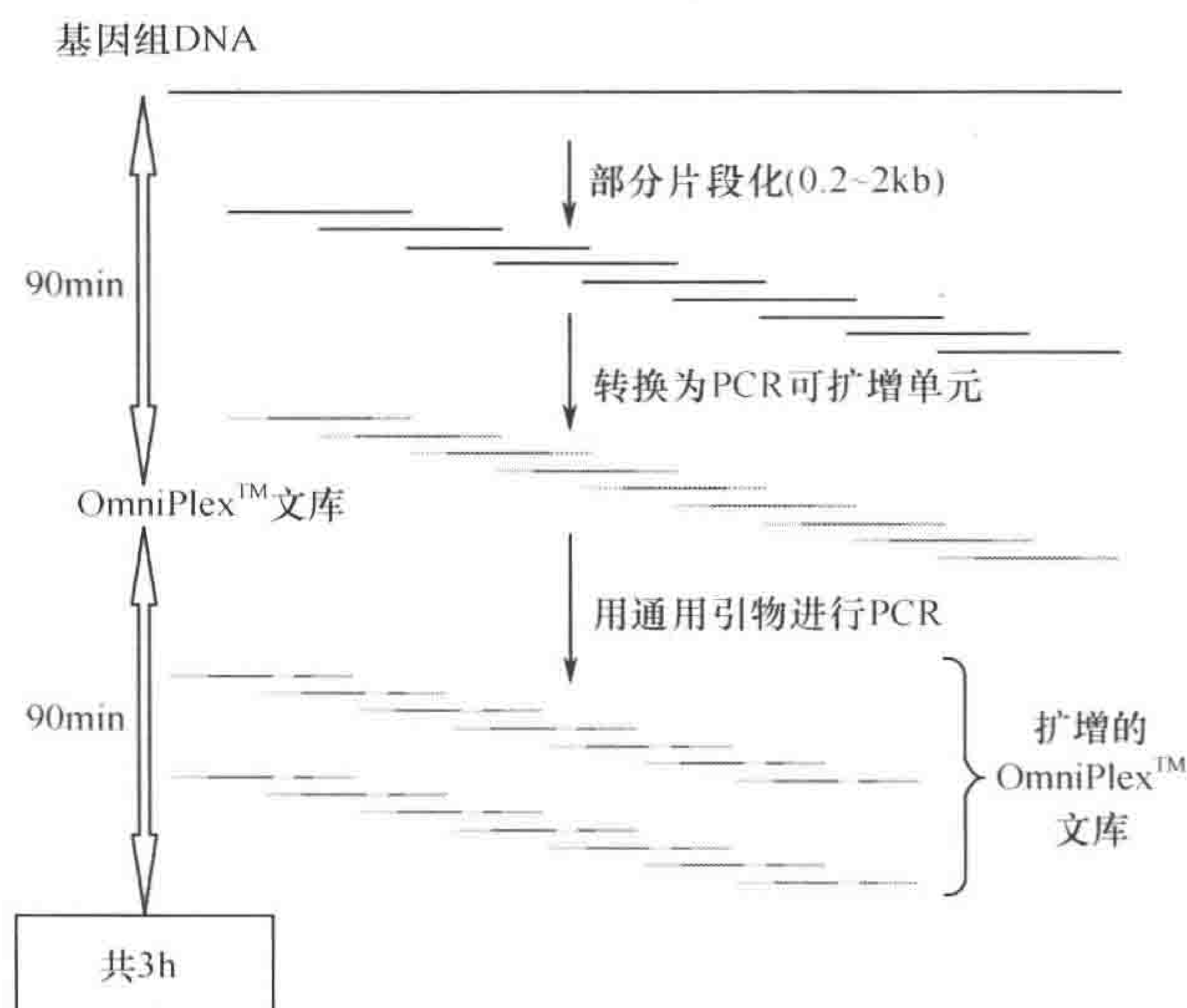



图 9-2 GenomePlex WGA 图示。GenomePlex 全基因组扩增分为三个主要步骤。首先，基因组 DNA 通过温和加热为基础的片段过程进行随机片段化（非序列依赖性）和变性。一个含 5′端通用前导区域的兼并引物延伸 DNA 结合退火。其次，进行引物延伸生成代表全基因组的 DNA 文库。最后，文库经有限循环次数的 PCR 典型性的扩增获得微克量 DNA 足够后续分析之用。对于组织（FFPE）或单个细胞样品，在上述步骤前还要进行起始基因组 DNA 的提取

后一种方法，作为 GenomePlex WGA 销售系列产品（sigma-aldrich），可获得 500～1000 倍无偏爱性扩增（图 9-3）。该平台对一些有问题样品的扩增已证明其在高精

度扩增多种不同的动植物组织的DNA是卓有成效的。但是，此方法最重要的价值是能够稳定地扩增降解的或数量非常有限的DNA（见本章实验方案二）。

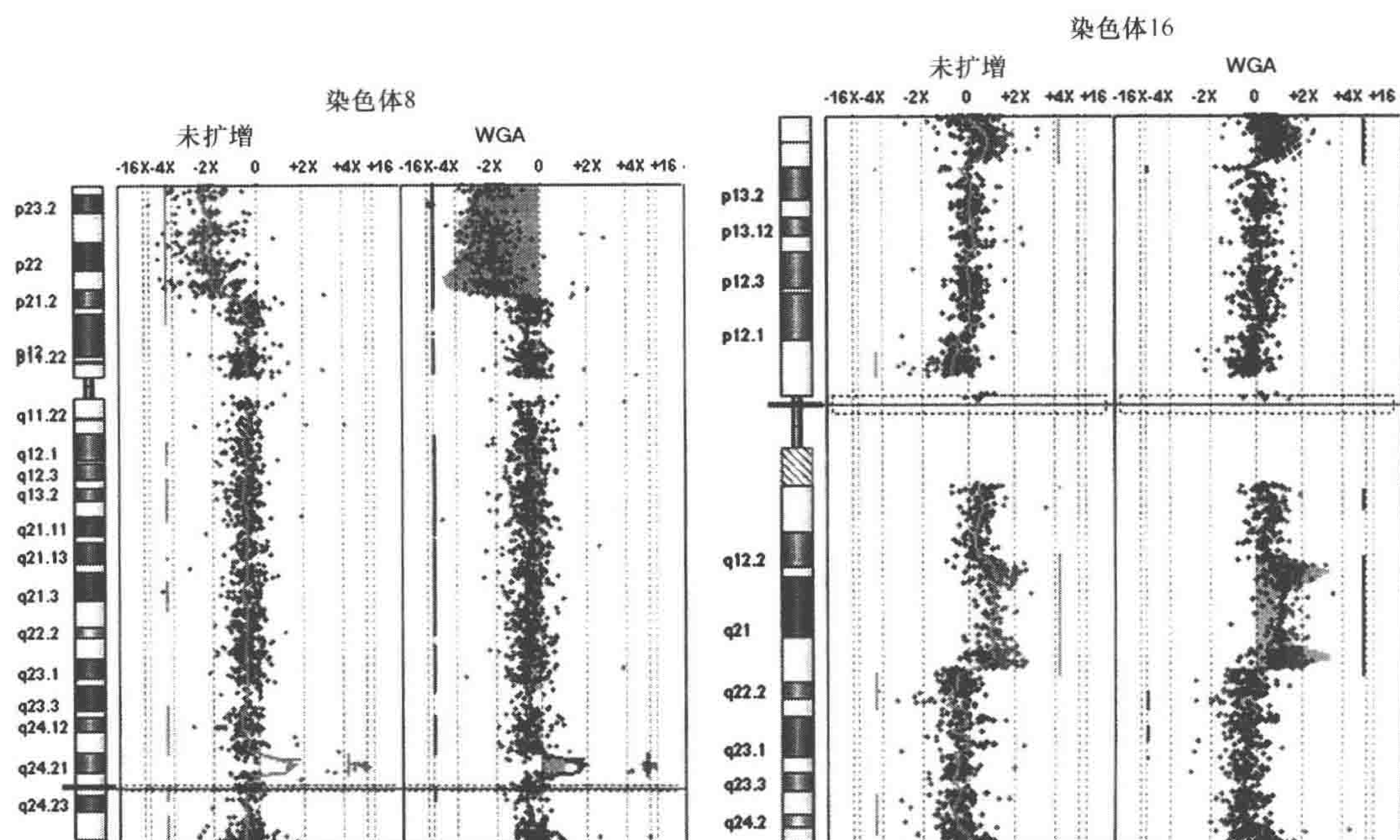



图9-3 比较基因组杂交（CGH）数据。化学处理或未处理的人类细胞在GenomePlex WGA程序中运行。Agilent 44K人CGH队列被用于比照WGA-扩增的物质（每个图片的右侧）和未经扩增的原初基因组DNA（图片左侧）。未扩增的基因组DNA和WGA DNA样品用BioPrime阵列CGH基因组标记系统分别标记为Cy3-和Cy5-。一个芯片表现了750 ng标记（Cy3和Cy5）的WGA DNA，第二个芯片显示的是2000 ng标记WGA DNA。这些阵列杂交40h。经漂洗，芯片在Agilent Scanner上进行分析。染色体8和染色体16在对比中获得代表性片段数据。WGA过程中没有产生与原初未扩增DNA对照可检测的偏差（Little et al. 2005）

本实验方案适用于扩增已预先提纯的不同来源的基因组DNA，可以是培养细胞、血液、唾液以及新鲜或冻存的组织。

# 材料

## 试剂

GenomePlex全基因组扩增试剂盒（sigma），包括

- 扩增母液混合物（10×）
- DNA片段化缓冲液（10×）
- 文库制备缓冲液
- 文库制备酶
- 文库稳定缓冲液

去核酸酶水

WGA DNA 聚合酶

来源于目标材料的纯化基因组 DNA

利用 GenEluteDNA 提取试剂盒（Sigma）提取的全血、培养细胞、植物组织和细菌 DNA 样品均可作为 WGA 的有效模板。从不同来源材料提取 DNA 的实验方案请参考本章其他部分，如培养细胞（方案一）、血液（方案五、六、七）、唾液（方案八）。关于植物 DNA 的提取另见第 7 章。

### 器材

移液器和枪头（专用）

分光光度计

PCR 仪（PE9700 或类似型号）

薄壁 PCR 管（0.2 mL 或 0.5 mL）或 96 孔 PCR 板

## 方法

### 片段化

（1）分离基因组 DNA 并通过紫外吸收（260 nm）确定浓度。准备 1ng/μL DNA 溶液。

（2）向含有 10 μL DNA（1ng/μL）PCR 管或 96 孔板中加入 1 μL 10×片段化缓冲液。

（3）将试管/96 孔板置于 95℃ 加热板或 PCR 仪 4 min 整。

该恒温过程对时间高度敏感，任何偏离都可能导致结果变化。

（4）迅速将样品冰浴冷却，短暂离心收集反应物（?）。

文库制备

（5）加入 2 μL 1×文库制备缓冲液。

（6）加入 1 μL 文库稳定缓冲液。

（7）彻底混匀，置于 95℃ PCR 仪中 2 min。

（8）迅速将样品冰浴冷却，离心加固，回置于冰浴中。

（9）加入 1 μL 文库制备酶液，涡旋振荡混匀，短暂离心。

（10）将样品置于 PCR 仪中，进行如下循环过程。

i. 16℃，20 min；

ii. 24℃，20 min；

iii. 37℃，20 min；

iv. 75℃，5 min；

v. 4℃保存。

（11）将样品从 PCR 仪中取出短暂离心。反应物可立即进行扩增，也可在－20℃保存 3 d。

扩增

(12) 向步骤 (11) 获得的 15 μL 反应物中加入以下试剂

7.5 μL 10×扩增母液混合物

4.5 μL 去核酸酶水

5 μL WGA DNA 聚合酶

(13) 彻底振荡混匀，短暂离心，开始 PCR。下列程序是对 PE9700PCR 仪或其他类似型号 PCR 仪的最优化方案。

| 循环数 | 变性 | 退火/延伸 |
|---|---|---|
| 第 1 循环 | 95℃ 3 min | |
| 14 循环 | 94℃ 15 s | 65℃ 5 min |

(14) 循环结束后，4℃维持反应或存于－20℃直至提纯及后续分析。推荐使用 Sigma's DNA cleanup 试剂盒进行 WGA DNA 提纯。WGA DNA 的稳定性与相同条件下保存的基因组 DNA 没有差别。在 DNA 定量前必须先提纯以除去未结合的引物，因为这些引物可能影响到扩增 DNA 后续应用。

## 疑难解答

问题：扩增后产量偏低

解决方案：

(1) 样品可能含 PCR 抑制剂或高缓冲溶盐。可用合适的微透析装置进行透析。该过程可稀释抑制成分。对透析产物进行定量，因为 DNA 在此过程中可能有损失。用 70%乙醇再提纯、沉降。

(2) 假如加入的 DNA 被严重降解或不足 10 ng，往往导致产量过低或最终产物不具典型性。添加多量的 DNA，这使得一些模板得到扩增。已有降解的 DNA 样品通过增加起始模板至 25～100 ng 而成功进行 WGA 扩增的例子。

(3) 可能需要更多的酶。当 DNA 聚合酶总量有限时 WGA 产量会受影响。每进行 75 μL 反应应添加至少 5 μL WGA DNA 聚合酶。这比增加反应的循环次数要好，因为后者可能导致 DNA 扩增终产量出现偏爱性。

(4) 反应后提纯可能有误。应使用保留单/双链 DNA 的方法。推荐使用 Sigma's DNA 纯化试剂盒。

(5) DNA 片段化反应可能太长或太短。4 min 片段化时间对各种 DNA 样品均能得到最理想结果。裂解不足或无片段化会导致产量过低或 WGA 最终产物缺少典型性。几乎所有条件下，10 minDNA 片段化也会造成产量过低，因为片段化过度导致 DNA 片断太小，不足以建立有效文库。

问题：qPCR 显示 WGA 引起目标基因表现明显偏爱性。

解决方案：

(1) 对照组可能不合适。基因组 DNA 只能在对照 DNA 剪切后才能与 GenomePlex

WGA 进行比较。使用经过片段化处理并合并的样品，或与经过水解剪切的 DNA 进行对比（Thorstenson et al. 1998）。

(2) DNA 样品量太少或过分降解。参见“低产量”问题解决方案的步骤（2）。

问题：阴性对照组（无模板）有产物。

解决方案：

一种或多种试剂可能沾染了外源 DNA。更换沾染的测试（尽管此问题可能不影响结果，但只有更换有问题的试剂才能得到干净的无模板对照）。

## 致谢

感谢 Shaukat Rangwala（密苏里，圣路易斯 MOgene）提供 Agilent 微阵列技术支持和疑难问题解答。

## 方案九　单细胞全基因组扩增

Chad Brueck, Clyde Brown, Steve Michalik, Deborah Vassar-Nieto, Ernie Mueller, and Gary Davis

*Sigma-Aldrich, St. Louis, Missouri 63103*

单个人类细胞，尽管只含有 6pg DNA，却可以清楚地研究细胞特异性调控和分化。揭示成熟、再生和遗传性疾病的未知诱因全都隐藏在多细胞生物体、有共同遗传来源的单个细胞中。细胞的重要性已经被一些研究所增强，曾经认为收集的混合细胞样本是同质性的，其实不然，这些混合细胞样品实际上常常由不同表型的细胞组成（Reyes et al. 1976）。

直到最近，全基因组扩增（WGA）的灵敏度不足以扩增单个细胞中 pg 量级的 DNA。为此，人们发明了 GenomePlex 单细胞全基因组扩增试剂盒突破这些限制。通过系统性优化 DOP 和 PEP 方法，灵敏度和基因组所含基因的代表性都得到了增强。虽然单细胞操作的失败率远远高于标准的 GenomePlex 操作，但它却提供了单细胞基因组百万倍级别的扩增，并允许在这种基础水平上作遗传分析。例如，图 9-4 显示了经过 GenomePlex 单细胞全基因组扩增试剂盒处理后的单个 U937 细胞产生的 SNP 数据。

## 材料

### 试剂

分离后的单个细胞

通过荧光激活细胞筛选技术（FACS）或激光捕获微分离（LCM）系统。不推荐稀释法。

GenomePlex 单细胞全基因组扩增试剂盒（Sigma），包括

扩增混合液（10×）

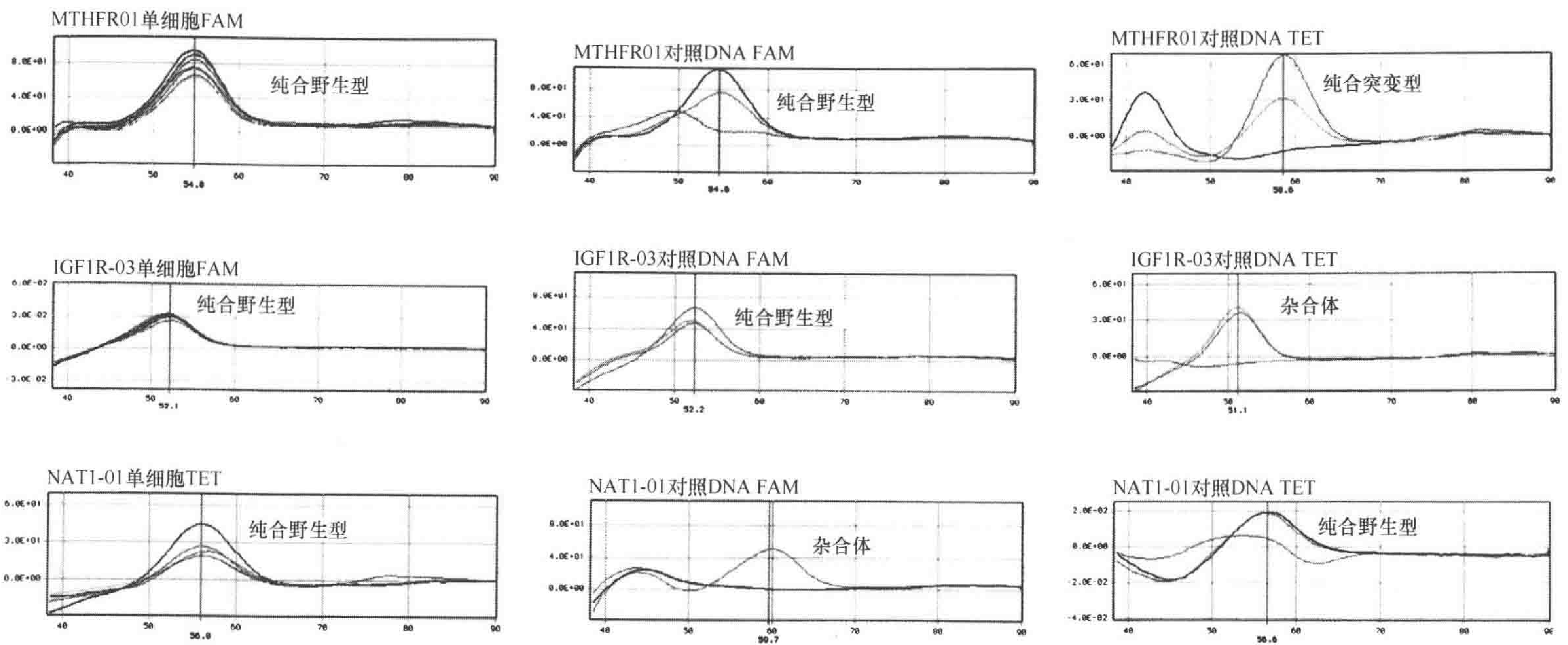



图9-4 单细胞SNP分析。利用单细胞全基因组扩增方法扩增人类白血病U937细胞。反应后产物通过GenElute PCR清除试剂盒进行纯化并通过分光光度计定量。之后利用ABI7700进行qPCR，通过MGB Eclipse SNP陈列对10ng单细胞全基因组扩增DNA进行分析。由于可产生后PCR解离曲线，MGB Eclipse探针有利于等位基因的辨别。解离曲线可确定SNP的熔点(Tm)。不同的等位基因具有差别很大的$T_m$值。FAM和TET荧光信号有助于区别野生纯合子、杂合子和突变纯合子。对全基因组扩增的单个细胞的DNA进行了三种SNP的检测：MTHFR01、IGF1R-03和NAT1-01。对未经扩增的阳性对照组与每个组都进行了评估。数据表明，与未经扩增的对照组相比，经过全基因组扩增的单个细胞的SNP得以保留(Barker et al. 2004；Gribble et al. 2004)

文库制备酶
文库稳定缓冲液
无核酸酶的水
蛋白酶 K 溶液
单细胞裂解和破碎缓冲液（10×）
单细胞文库制备缓冲液
全基因组扩增聚合酶

### 器材

移液器和枪头（专用）
分光光度计
PCR 仪（PE9700 或类似型号）
薄壁 PCR 试管（0.2 mL 或 0.5 mL），或 96 孔 PCR 板

## 方法

### 裂解和破碎

（1）利用 FACS、LCM 或其他方法将单个细胞分离至一个准备进行 PCR 的容器中。如果细胞要分类筛选，缓冲液应该具有较低的离子浓度，如 Tris EDTA（TE）缓冲液，而且应使用最小的筛选体积。

（2）向单细胞样品中加入足量体积的水至 9 μL。

（3）向 32 μL 10×单细胞裂解破碎缓冲液中加入 2 μL 蛋白酶 K 溶液，配制成实验用裂解破碎缓冲溶液。充分搅拌混匀。

（4）向单细胞样品中加入 1 μL 步骤（3）中新配制的裂解破碎缓冲液。充分混匀。

（5）50℃温育 1 h，然后加热至 99℃，准确地加热 4 min。冰上冷却，在进行文库制备前要将样品离心。

此步骤温育对时间非常敏感。任何偏差都可能使结果发生变化。

### 文库制备

（6）加入 2 μL 1×单细胞文库制备缓冲液。

（7）加入 1 μL 文库稳定溶液。

（8）充分混匀并在 PCR 仪上加热至 95℃，持续 2 min。

（9）立即在冰上冷却样品，通过离心使样品稳定，再转移到冰上。

（10）加入 1 μL 文库制备酶液，充分搅拌，短暂离心。

（11）将样品置于 PCR 仪中，按如下步骤温育。

i. 16℃，20 min

ii. 24℃，20 min

iii. 37℃，20 min

iv. 75℃，5 min

v. 4℃ 保温

(12) 将样品从 PCR 仪中取出并短暂离心。样品可以立即进行扩增或在−20℃保存3 d。

### 扩增

(13) 向步骤（12）的全部 14 μL 反应体系中加入以下试剂。

7.5 μL 10×扩增混合液

48.5 μL 无核酸酶的水

5 μL 全基因组扩增 DNA 聚合酶

(14) 充分搅拌，短暂离心，开始 PCR 循环。下面是为 PE9700 或同等 PCR 仪优化的程序。

| 循环次数 | 变性时间 | 复性/延伸时间 |
|---|---|---|
| 第 1 次循环 | 95℃，3 min | |
| 25 次循环 | 94℃，30 s | 65℃，5 min |

(15) 循环完成后，将反应体系保持在 4℃或在纯化和分析使用前以−20℃保存。推荐利用 Sigma 的 PCR 纯化试剂盒进行全基因组扩增 DNA 的纯化。全基因组扩增 DNA 的稳定性与同样条件下保存的基因组 DNA 相同。在 DNA 定量前，样品必须进行纯化以除去未结合的引物，否则会干扰一些下游操作。

## 疑难解答

问题：PCR 产量不高。

解决方法：

(1) 样品可能含有盐或抑制剂。改变筛选步骤，使抑制剂或盐不被带入。

(2) 模板质量太差。单个细胞中的 DNA 可能在分离过程中被分解。单个细胞的储藏可能不适当。用多于一份的样品进行 WGA，这有助于从根本上排除随机的质量问题，但如果筛选过程损害了细胞，则无助于解决问题。

(3) 反应后的纯化可能不合适。使用保持单链和双链 DNA 的方法。推荐使用 Sigma 的 PCR 纯化试剂盒。

(4) 可能没有捕获到单个细胞。确保在准备进行 PCR 的试管中有单个细胞。同时，在加入单细胞裂解/破碎缓冲液后充分搅拌混匀。

(5) 如果将样品稀释至单个细胞，要使用多组反应物，因为试管有可能是空的。

问题：定量 PCR 显示出全基因组扩增对目的基因的再现有显著的偏爱性。

解决方法：

(1) 对照组可能不合适。只有当对照组 DNA 被剪切后，基因组 DNA 才能与全基因组扩增 DNA 进行比较。使用那些破碎实验方案获得的混合样品，或者与经水解剪切

过的 DNA 作为对照进行比较（Thorstenson et al. 1998）。

（2）DNA 样品量有限或者被过度降解。见低产量问题的解决方法。

问题：阴性对照组（无模板）获得产物。

解决方法：

可能一种或多种试剂被外源 DNA 污染。替换受影响的成分（尽管这一问题可能不影响结果，但只有通过替换受影响的成分才能获得一个清洁的无模板的对照组）。

## 感谢

我们感谢 Michael Speicher（人类遗传研究所，Tu Munich）的 WGA 程序的 β-检验，Shaukat Rangwala（Mogene，St. Louis，Missouri）的 Agilent 微阵列和技术的疑难解答，以及 Barbara Pilas（UIUC，Urbana-Champaign，Illinois）和 Joy Eslick（SLU，St. Louis，Missouri）两人所做 FACS 的单细胞全基因组扩增。

## 方案十　使用 Φ29 DNA 聚合酶进行全基因组扩增

Noël P. Burtt

*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139*

DNA 样品的品质和含量是所有遗传分析的基础。DNA 是一种珍贵的有限资源，在人类疾病研究中，样品 DNA 的获取常常被分离方法和人的来源所限制。另外，没有完整的样品 DNA，法医分析和考古研究通常无法进行。所以，保存或者增加 DNA 储备量的方法对这些研究的成功起关键作用。以往，为保存或维持 DNA 储备，发明了耗费成本高和劳动强度大的 Epstein-Barr 病毒转化细胞系。细胞系的创立除了创造一种可更新的 DNA 资源外，在许多方面也很有价值的；但是创立细胞系所要付出的成本和努力以及创建初始对细胞完整性的要求，限制了这一方法的效用。近来，人们希望通过几种统一而健全的方法扩增全基因组，同时限制消耗成本和时间。两种常见的以 PCR 为基础的方法是简并寡核苷酸引物 PCR（DOP）（Telenius et al. 1992）以及引物延伸预扩增（PEP）（Zhang et al. 1992）。这些方法有一定作用。然而，PCR 步骤会引入人工产物和不均一的基因组区域，可能导致扩增产物中出现短片段（Lovmar et al. 2003）。

最近，利用 Φ29 DNA 聚合酶独特性质的全基因组扩增，已经被多次应用于产生稳定的高可信度基因组拷贝（如 Hoson et al. 2003；Luthra and Medeiros 2004；Paez et al. 2004；Pask et al. 2004；Bergen et al. 2005）。利用 Φ29 DNA 聚合酶的全基因组扩增通过多重链替代以及随后在随机引物上的滚环复制，产生无偏差的基因组重现（Dean et al. 2001）。这种原先被用于质粒、细菌噬菌体和其他环状基因组滚环复制的酶被移植到线形基因组，利用它独特的链替代行为（图 9-5）。伴随着随机寡核苷酸六聚体结构的加入，基因组的 10 000 倍拷贝可以产生比前述方法大得多的片段尺寸（10kb 或更大）。利用商业出售的试剂盒，通常可以从少至 10 ng 的反应物 DNA 获得 35～50 μg 产物。高质量的初始 DNA 是成功的关键，因为被分解或损害的 DNA 不会产生健全的产物。这种方法的主要优点是不需要热循环（确切地讲，反应是恒温的）；酶

的错误率比较低，每百万碱基少于 10 个的错误水平（大约与 PCR 相同）；以及可得到相对均衡的基因组重现。

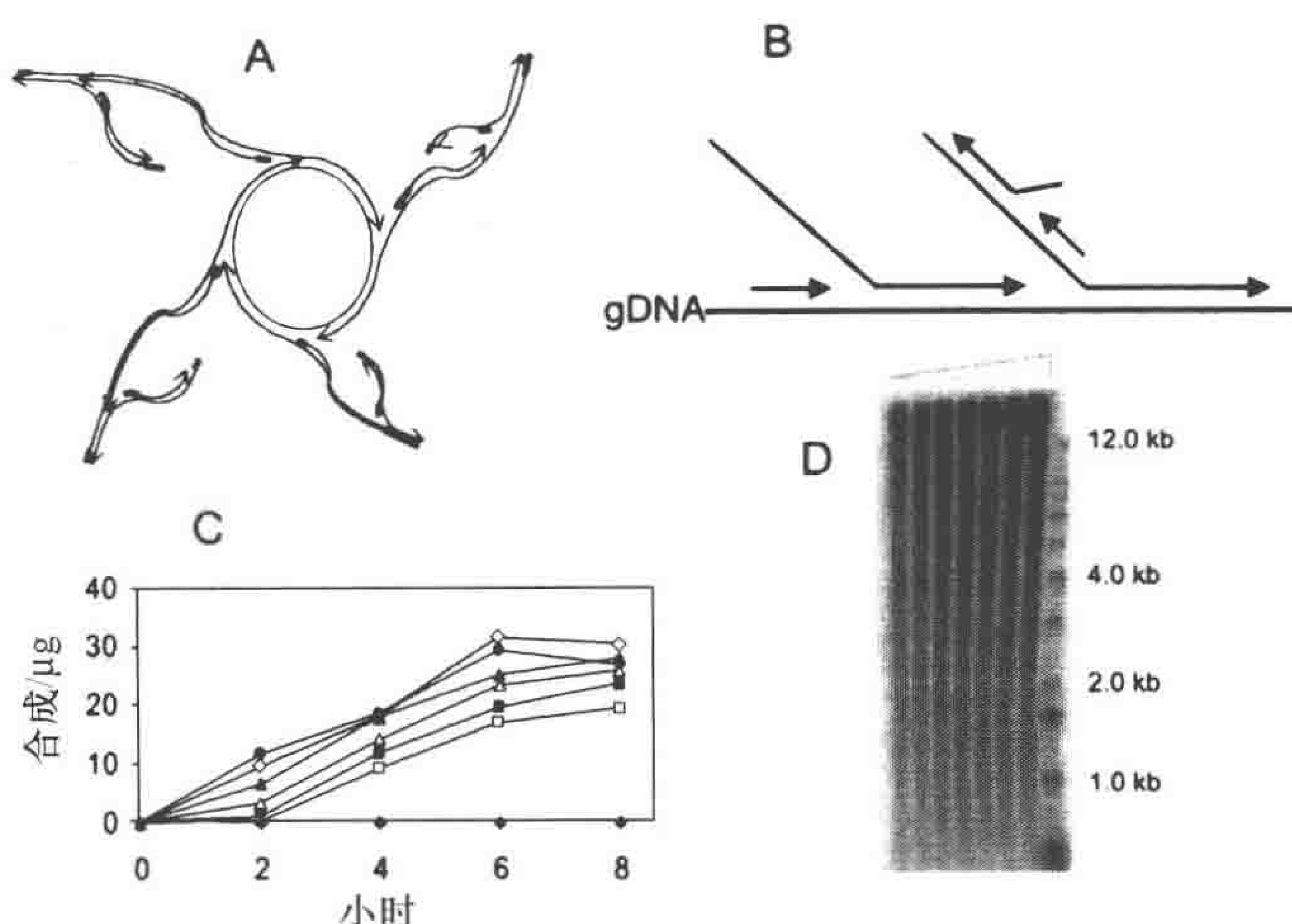



图 9-5　多重替代扩增（MDA）策略和产物特征鉴定。A. 环状 DNA 的随机引物滚环复制。DNA 合成由随机寡核苷酸引物引发。3′端由箭头表示。加粗区域表示引物。第二次引物配对事件发生在被替代的产物 DNA 链上。B. 基因组 DNA 的 MDA 图示。第二次引物配对事件由初始产物引发。C. 模板浓度对扩增产量的影响。总量 100fg～100 ng 人类基因组 DNA 在 30℃下通过 MDA 进行扩增。各等份是在指定的量化 DNA 合成的时间内从单一反应体系中取得。符号：(●) 10 ng 基因组 DNA 模板；(○) 1ng；(▲) 100pg；(△) 10pg；(■) 1pg；(□) 100fg；(◆) 无引物。D. 扩增产物大小的变性凝胶条带分析。在 C 中展示的放射性标记的扩增产物通过碱性琼脂糖凝胶（1%）进行电泳，烘干的凝胶被暴露在荧光屏下成像。反应产物以递增的 DNA 模板量为序进行加样，如凝胶上的图像所示

尽管这种方法总体上比较健全，但必须小心，最大限度减少样品的交叉污染，因为反应体系中的任何 DNA 都会被扩增。另外，在遗传分析中利用扩增产物之前，质量控制检验是非常关键的。除了通常的质量控制步骤外，还要采取一些措施，如 PicoGreen 检测以确定 DNA 的量。出售商建议进行基于 PCR 的 TaqMan 检测以确认扩增是否发生以及是否是人类特异性的 DNA（而不是六聚体或污染的 DNA 的随机扩增，这可能发生在没有合适的人类 DNA 作为模板的反应体系中）。另一项有价值的建议是检查在扩增产物和原来未扩增 DNA 上的遗传标记的信息模式的基因型。如果原来的材料可用的话，这显然是可性的（我们估计在一个多元反应体系中要对约 50 个 SNP 进行基因分型需要 40 ng DNA）。另外，还可以检验两个产物的基因型以确保样品保持一致性。

## 材料

注意：带＜!＞标记的试剂请按附录指导小心操作

## 试剂

由于Φ29 DNA聚合酶使用的专利限制，这个实验过程及试剂是按照两种商业用WGA试剂盒的描述。

酸性缓冲液（pH 4）9（B溶液；只用于REPLI-g试剂盒）

扩增溶液（REPLI-g试剂盒中的REPLI-g缓冲液，4×；GenomiPhi试剂盒中的反应缓冲液）

下列是两种试剂盒中共有的组分

37 mmol/L Tris-HCl（pH7.5）

50 mmol/L KCl

10 mmol/L $MgCl_2$

5 mmol/L $(NH_4)_2SO_4$ <!>

1 mmol/L dNTP

50 μmol/L 抗外切酶核酸六聚体

1 单位/mL 酵母焦磷酸酶

Φ29 DNA聚合酶

该酶用于两种试剂盒中。－80℃储存。在滚环扩增中使用这种酶的权利属于GE Healthcare。

EDTA（0.5 mol/L，pH 8）

基因组双链DNA（100 ng）

在无核酸水中稀释。以TE稀释的反应物DNA会减少反应产量。

KOH（5 mol/L）

无核酸酶的水

PicoGreen

0.5× TE缓冲液（200 mmol/L Tris，20 mmol/L EDTA，pH 7.5）

WGA试剂盒［REPLI-g试剂盒（Qiagen）或GenomiPhi试剂盒（GE Healthcare）］

## 器材

冰

Macherey-Nagel NucleoFast板［可选；步骤（17）］

微量离心管（用于单个反应）

特殊仪器（用于96孔平板中的100组反应）

96 MultiMek机器人和升降台（如Beckman Coulter Biomek FX-LP-8927A），当需要大容量处理多于500个样品时使用。

PCR板（96孔，如ABgene Thermo-Fast 96 Skirted、Low Profile-AB-0800）

PCR仪（如Applied Biosystems Dual 96-Well GeneAmp PCR System 9700-4343176或同等型号）

在 REPLI-g 试剂盒手册中，加热设备没有特别说明。可使用水浴或培养箱作为替代品。

滴定盘摇床（如 Barnstead/Lab-Line 57019-600 或同等型号）

试管（圆锥形）

涡旋振荡混合器

## 方法

反应的开始时间取决于反应所使用的仪器。如果使用 PCR 仪，给机器设定程序冷却至 4℃。反应混合物可以在若干小时内保持这一温度。如果使用培养箱，下午开始反应，这样温育可以在第二天早上结束。虽然这一方案基于 REPLI-g 试剂盒，但我们将参考 GenomiPhi 试剂盒以指出两种方案的不同之处。

### 准备

（1）在室温下解冻 B 溶液，将扩增溶液和 DNA 聚合酶置于冰上。

（2）向微量离心管中加入 2.5 μL 基因组 DNA。对于 100 组反应，向一个 0.2 mL 96 孔有沿板中加入 2.5 μL 基因组 DNA。

要获得理想产量，需要使用 50 ng/μL 浓度的基因组 DNA。如果推荐的量（100 ng）无法达到，将实验方案相应地调整为适于低浓度（大于 20 ng/μL）反应底物基因组 DNA。

### 变性

这一步是分开基因组 DNA 双链，提供单链模板以促进随机引物的结合。

（3）按如下方法制备变性缓冲液（A 溶液）。

| | 单个反应 | 100 组反应 |
|---|---|---|
| 0.5 mol/L EDTA | 1 μL | 100 μL |
| 5 mol/L KOH | 4 μL | 400 μL |
| 无核酸酶的水 | 35 μL | 3500 μL |
| 低反应物 DNA | | |
| | 单个反应 | 100 组反应 |
| 200 mol/L KOH | 5 μL | 500 μL |
| 无核酸酶的水 | 35 μL | 3500 μL |

（4）向 DNA 样品中加入 2.5 μL A 溶液，用移液器混合均匀。

对于 100 组反应，当溶液加入最后一个样品之后就开始温育阶段。

（5）在室温下温育 2.5～3 min。

在 GenomiPhi 方案中，向 DNA 中加入 22.5 μL 样品缓冲液。将样品加热至 95℃，保持 3 min，然后使用 PCR 仪或者将样品置于冰上，使样品降温至 4℃。REPLI-g 试剂

盒不推荐加热，因为加热会使 DNA 变性过度，并导致断裂（Dean et al. 2001）。

## 中和

只要求在 REPLI-g 方案中使用，以停止变性并使样品 pH 降低至适合酶活性的环境。

（6）以 1∶10 的比例稀释 B 溶液（1.8 mL B 溶液，16.2 mL 无核酸酶的水）。每次实验要新配制该溶液。

一旦 B 溶液解冻并处于室温，在反应后将不能再次使用。要保证均一的混合物，将 B 溶液全部稀释。

（7）向 DNA 样品中加入 5 μL B 溶液，用移液器混合均匀。

## 扩增

将随机引物、核苷酸六聚体和酶混合，引发 MDA 的滚环扩增过程。

（8）按如下方法，从表中找出合适的体积，在冰上制备混合母液：

i. 在一个圆锥形试管中混合 REPLI-g 缓冲液（4×）和水，搅拌混匀。

ii. 加入 REPLI-g DNA 聚合酶，再次搅拌混匀溶液。

| | 1 组反应 | 100 组反应 |
|---|---|---|
| REPLI-g 缓冲液(4×) | 12.5 μL | 1.25 mL |
| REPLI-g DNA 聚合酶 | 0.5 μL | 0.05 mL |
| 无核酸酶的水 | 27.0 μL | 2.7 mL |
| 总计 | 40.0 μL | 4.0 mL |

（9）向 DNA 样品中加入 40 μL 混合母液，不要涡旋搅拌。

溶液比较黏稠，太多的搅动将引起泡沫。在 GenomiPhi 方案中，向 DNA 样品中加入酶混合物以及反应缓冲液。

## 温育

这一步最好用 PCR 仪进行，因为温度要维持恒温 30℃，而且在 PCR 仪中进行多组样品的反应比在培养箱中更加均一。引物配对和扩增在基因组中平均地发生，而且不会产生偏爱性。

（10）在 PCR 仪中 30℃温育 16 h。

在 GenomiPhi 方案中，只温育 4 h（因为反应缓冲液中引物和 dNTP 的量与 REPLI-g 混合物不同）

（11）在温育阶段末，将 DNA 样品加热至 65℃，保持 3 min，然后使样品降温至 4℃。

温度峰值中的 3 min 对阻止 Φ29 DNA 聚合酶的外切酶活性很有必要，否则酶将开始降解产物。

## 稀释

这一步骤使全基因组扩增产物变得均一，由于反应体系的初始黏稠度，这一步骤是

必需的。稀释后的工作浓度将为 125ng/μL。

(12) 反应一结束，立即以 1∶10 比例在 450 μL 0.5× TE 缓冲液中稀释产物。对长期储存来说，这样比在水中稀释更好。

(13) 将 DNA 样品置于滴定盘摇床上，以温和的速度在室温下过夜。

如果全基因组扩增产物未被稀释，沉淀物的再降解将极为困难。同样，未被稀释样品的定量（如用 PicoGreen）将导致高度波动的浓度。

### 产物的定量

(14) 将全基因组扩增的储备产物进行 1∶10 稀释。如果全基因组扩增反应获得足够产物，稀释后的产物浓度应为 12.5ng/μL。

(15) 在这一浓度范围，可使用 PicoGreen（自动且一致）估算产量（见疑难解答）。

### 可选择的质量控制检测

(16) 使用 TaqMan 拷贝数检测，在基因组两个容易被分解的位点评估 DNA 拷贝数（Hosono et al. 2003）。

(17) 纯化全基因组扩增产物。真空过滤是一种保守的措施，除去多余的引物和六聚体。Macherey-Nagel NucleoFast 板使用超滤以除去溶液中所有的盐和小于 10 个核苷酸的引物。

如果能够负担纯化的成本，而且 DNA 将被用于下游基因型分析，就必须进行纯化。纯化将除去任何多余的可能干扰下游反应的 PCR 引物和核苷酸六聚体等。

## 疑难解答

问题［步骤（15）］：产物的浓度被估算过高。

解决方法：

PicoGreen 是定量的最佳选择。测量 OD 值并无效，因为它含高估 DNA 的量并给出高于预期的浓度。类似地，定量仪法也会高估产物浓度，因为残余的六聚体和引物额外增加了量产物。另外，这一程序并不是自动化的。

问题［步骤（15）］：PicoGreen 产生了假阳性值。

解决方法：

没有 DNA 模板的对照反应产物将会显示假阳性的 PicoGreen 值。剩余的双链核苷酸六聚体将会使 PicoGreen 产生与含有 DNA 模板的样品一样结果。如果可能，使用另外的验证方法；特别要使用一种能显示人类扩增 DNA 存在的方法，如基于 PCR 的 TaqMan 检测。

## 方案十一 利用 Chelex 从法医样品中提取 DNA

John M. Butler

*Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8311*

这个方案描述了使用螯合离子交换树脂悬液，即 Chelex 从法医样品中提取 DNA 的过程。这是由武装部队 DNA 鉴定实验室（Rockville，Maryland）的标准操作步骤改编而来的方案。

## 材料

注意：请参考附录对带有<!>标记的试剂小心操作。

### 试剂

Chelex 溶液（20%：2 g Chelex 溶于 10 mL 水；5%：0.5g Chelex 溶解于 10 mL 水）

二硫苏糖醇（DTT；1 mol/L）<!>

提取缓冲液（10 mmol/L Tris，pH 8.0；100 mmol/L NaCl；50 mmol/L EDTA，pH8.0；0.5%SDS<!>）

蛋白酶 K（20 mg/mL）<!>

样品（棉签或碎衣物）

### 器材

冰存管（无菌，带标签）

镊子［可选；见步骤（6)］

微量离心管（1.7 mL）

移液器枪头

无菌剪刀或手术刀片

离心收集管：Spin-EASE 试管

如果没有 Spin-EASE，用套着底部有刺穿小孔的 0.5 mL 薄壁试管的（与 Spin-EASE 的吊篮部分相当）放在 1.7 mL 试管中（与 Spin-EASE 的下部相当）替代（背负式旋转）

涡旋搅拌混合器

水浴锅（37℃和沸腾）

称重盒

# 方法

## 样品中上皮细胞和精细胞的回收

（1）将合适数量的 Spin-EASE 试管的下部和 1.7 mL 微量离心管贴上特殊的标签。

（2）向每个有标签的 Spin-EASE 提取管中加入 800 μL 无菌水。在这时设置底物对照组（如果可用的话）和试剂空白组。

（3）用干净的剪刀或无菌手术刀片，在洁净的台面上分离棉签或碎布。向试管中加入样品（1/2 的棉签或约 3 mm 见方的碎布）。

（4）快速涡旋搅拌，在室温下保持 30 min。

（5）涡旋搅拌 10 s 以从底物中搅起细胞。脉冲离心。

（6）用镊子或枪头将底物转移到 Spin-EASE 试管的吊篮。将含有样品底物的吊篮放进 Spin-EASE 体系的下部。

（7）样品在微量离心管中以最高速（10 000～15 000 r/min）离心 1 min。

（8）给 Spin-EASE 的吊篮贴上标签，并将其放入通风橱中的一个称重盒里。让样品在空气中晾干过夜。当样品晾干时，将它重新包装并作为证据储存。

（9）在不触动沉淀物的情况下，倒出上清液，只保留 50 μL。通过移液器枪头搅拌将细胞碎屑打散。

这些沉淀物中含有上皮细胞和精细胞。

## 细胞的裂解及 DNA 的提取

（10）向重悬的细胞碎屑中加入 150 μL 无菌水和 2 μL 蛋白酶 K（20 mg/mL）溶液，轻轻混匀。

（11）37℃温育约 1 h 以裂解上皮细胞，但不要超过 2 小时以尽量减少精细胞的裂解。

（12）将微量离心管以 10 000～15 000 r/min 离心 5 min。

（13）在不触动精细胞沉淀物的情况下，将 150 μL 上清液转移到一个新的 1.7 mL 微量离心管中，标上“E”（上皮，或雌性部分）。加入 50 μL 20% Chelex 溶液。室温下保存直到步骤（17）。

（14）在 500 μL 提取缓冲液中将沉淀物打散重悬。短暂涡旋搅拌，将微量离心管以 10 000～15 000 r/min 离心 5 min。倒出上清液，只保留 50 μL。重复两次，总计洗涤 3 次。

（15）在最后一次缓冲液洗涤后，向沉淀物中加入 1 mL 无菌水。短暂涡旋搅拌并在微量离心管中以 10 000 r/min 离心 5 min。倒出上清液，只保留 50 μL。

（16）向精细胞沉淀物加入 150 μL 5% Chelex，2 μL 蛋白酶 K（20 mg/mL）溶液和 7 μL 1 mol/L DTT，轻轻混匀。这是污物痕迹的雄性部分，或精子。把它标上“S”。

（17）37℃温育 30～60 min。

（18）以高速涡旋振荡混合上皮细胞样品（E）和精细胞样品（S）5～10 s。

(19) 将微量离心管以 10 000～15 000 r/min 离心 10～20 s。

(20) 在沸水浴中保持样品 8 min。

(21) 涡旋搅拌试管 5～10 s。

(22) 在微量离心管中以 10 000～15 000 r/min 离心样品 3 min。

(23) 对样品定量（见方案十二）。然后，扩增样品。

所有核心的犯罪证据提取物必须在扩增前定量。

(24) 放置 2～8℃，Chelex 珠子可以短期储存。要使用的 Chelex 珠子短期储存后，重复步骤 (20)（可选）、(21) 和 (22)。对于长期储存，将上清液从 Chelex 珠子上转移到一个无菌的有标签试管中，至少－20℃冰冻。

## 方案十二　在多元 PCR 扩增前对法医样品中的 DNA 浓度进行估算

John M. Butler

*Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8311*

这一实验方案描述了在 PCR 扩增前使用定量 PCR 对法医样品中提取的 DNA 进行定量。这是由武装部队 DNA 鉴定实验室（Rockville，Maryland）的标准操作步骤改编而来的方案。

### 材料

#### 试剂

Quantifiler qPCR 试剂盒（Applied Biosystems），包括：

Quantifiler 人类 DNA 标样

Quantifiler 人类引物混合物

Quantifiler PCR 反应混合物

参见 Quantifiler 人类 DNA 和 Y 人类男性 DNA 定量试剂盒使用手册（Applied Biosystems 2004）以获得更多信息。

样品（按实验方案十一制备）

TE 缓冲液

#### 器材

ABI Prism 7500 序列检测系统（SDS）（实时 PCR）

光学反应盘（96 孔）

移液器和枪头

测试试管

涡旋搅拌混合器

# 方法

## 标准样、对照和样品的制备

(1) 启动 ABI Prism 7500 序列检测系统。

(2) 当天新制备的 Quantifiler 人类 DNA 标准。将 8 支试管分别标注 1～8、日期和首字母。向每支试管加入适量 TE（见下表）并涡旋搅拌 Quantifiler 人类 DNA 标准以混合均匀。将 5 μL DNA 标准液转移到试管 1，用移液器来回抽吸，然后搅拌 30 s 以混合均匀。换一个枪头，从试管 1 中移走 5 μL，将其加入含有 10 μL TE 的试管 2。按上述方法混合。重复上述步骤直到试管 7 的 5 μL DNA 被加到试管 8。

确保当用移液器 DNA 加入试管时来回抽吸数次，从一个试管向另一个试管加入 DNA 后要更换枪头，在进行下次稀释前涡旋搅拌以混合 DNA/TE 溶液。

| 标准 | 浓度/(ng/μL) | 稀释组合 | 稀释因子 |
|---|---|---|---|
| 标准 1 | 50.000 | 5 μL[200 ng/μL 储备]+25 μL TE 缓冲液 | 4× |
| 标准 2 | 16.700 | 5 μL[标准 1]+10 μL TE 缓冲液 | 3× |
| 标准 3 | 5.560 | 5 μL[标准 2]+10 μL TE 缓冲液 | 3× |
| 标准 4 | 1.850 | 5 μL[标准 3]+10 μL TE 缓冲液 | 3× |
| 标准 5 | 0.620 | 5 μL[标准 4]+10 μL TE 缓冲液 | 3× |
| 标准 6 | 0.210 | 5 μL[标准 5]+10 μL TE 缓冲液 | 3× |
| 标准 7 | 0.068 | 5 μL[标准 6]+10 μL TE 缓冲液 | 3× |
| 标准 8 | 0.023 | 5 μL[标准 7]+10 μL TE 缓冲液 | 3× |

(3) 将一试管 Quantifiler 人类引物混合物从－20℃冰箱中取出，解冻，涡旋搅拌均匀并短暂离心。从冰箱中取出 Quantifiler PCR 反应混合物。用移液器来回抽吸混合，但不要涡旋搅拌，因为这将产生泡沫，从而影响定量检测。

(4) 在一张工作记录单上按照加入平板的顺序记录标准和样品。一个平板最多可用于 76 个样品。

(5) 涡旋搅拌并快速离心所要定量的标准和样品。

(6) 按下述方法制备样品混合母液。

混合液

N＝将要扩增的样品数量

N＝ ____________

成分

Quantifiler 人类引物混合物（N＋4）×10.5 μL

Quantifiler PCR 反应混合母液（N＋4）×12.5 μL

(7) 用移液器将样品来回抽吸以均匀混合。用移液器将 23 μL 样品混合液移入 96 孔光学反应盘的孔中。

(8) 保留 A2～B2 孔和 G11～H11 孔为 4 个无模板对照组（NTC），并为 DNA 标

准样保留 A1～H1 和 A12～H12 16 个孔。

(9) 对于无模板对照组 (NTC)，用移液器向指定的孔中加入 2 μL 无菌水。

(10) 涡旋搅拌 DNA 标准样 15 s，沉淀，用移液器向设定的孔中加入 2 μL 适量的标准样 DNA 液(50 ng/μL、16.7 ng/μL、5.56 ng/μL、1.85 ng/μL、0.620 ng/μL、0.210 ng/μL、0.068 ng/μL、0.023 ng/μL，按步骤 (2) 中的表)。

(11) 对于所有其他样品，用移液器向设定的孔中加入 2 μL 样品。

(12) 用 ABI Prism 光学反应薄膜遮住反应盘，短暂涡旋搅拌，然后以 1000 r/min 离心 30 s。

(13) 把 ABI 压缩式防震垫放在光学反应盘上面，并把反应盘放在 7500 检测系统中 (确保反应盘被定向，这样 A1 孔位于反应盘的左后方)。

## 检测

(14) 打开 Sequence Detection System 软件。选择“File>New”。在对话框中，“Assay”应被设定为“Absolute Quantitation”，“Container”应被设定为“96-Well Clear”，而“Template”应被设定为“Blank Document”。点击“OK”关闭对话框，打开一个新的 plate 文件。

(15) 选择“Tools>Detector Manager”，点击时按住“CTRL”键选择“Quantifiler Human”和内部 PCR 对照组 (IPC) 探测器。选择“Add to plate document”以添加到反应盘然后选择“Done”以退出“Detector Manager”。

(16) 按如下方法应用探测器制定标准：

i. 选择“View>Well Inspector”以打开对话框。这将显示出探测器已经被添加到反应盘。确认选择了 ROX 作为“Passive Reference”。

ii. 在 Plate 表中，选择对应一个特定的定量标准样的孔 (如果选择多个孔可以按住“Control”键的同时点击多个孔)，然后返回“Well Inspector”。

iii. 对可应用的探测器、IPC 和 QuantifilerHuman 选择“Use”框。

iv. 对 IPC，在“Task”列中选择“Unknown”。

v. 对 Quantifiler Human，从“Task”列的下拉列表中选择“Standard”，然后为合适的探测器选择“Quantifiler”区域。输入孔中 DNA 的量。

vi. 输入“Sample Name”(如标准 1、标准 2)。

(17) 按如下方法应用探测器检测未知样品

i. 在 Plate 表中，选择对应所有未知样品的孔然后返回“Well Inspector”。

ii. 对已选的孔，选择“View>Well Inspector”，然后对可应用的探测器、Quantifiler Human 和 IPC 进行“Use”框检查。确认选择了 ROX 作为“Passive Reference”。

iii. 在“Well Inspector”打开的情况下，点选每个孔并输入“Sample Name”。未使用的孔应该被指定为“Not In Use”。

(18) 完成后，点击仪器上的键盘输入程序。

i. 删除“Stage 1”保持步骤。扩增程序应为如下样式

保持：95℃，10 min (以活化 Taq 聚合酶)

40 个循环：95℃，15 s

60℃，1 min

ii. 输入体积为 25 μL。

iii. 检查确认框中设定为 9600 Emulation。

iv. 将文件存为 SDS 文件（* .sds）。选择开始。

（19）当程序完成，从“Analysis”中选择“Analysis Settings”确定分析设置。将探测器指定为“All”。阈值应显示为 0.2，基线开始于 6 个循环，底线终止于 15 个循环。点击“OK”。从“Analysis Menu”中选择“Analyze”。

（20）从“File”菜单中选择“Export”，并输出结果。结果将以微软 Excel 工作表的形式出现。将“Baseline StdDev”，“delta Rn”和“Quantity”等列设定为四位小数。在“Quantity”列后再加入一列，命名为“Concentration（ng/μL）”。把所有的定量结果都除以因数 2，在“Sample Placement and Results”工作表中以 ng/μL 记录定量值。打印出样品位置和 Excel 工作表并放在合适的文件夹中。

（21）从探测器下拉列表中选择“IPC”。如果 IPC 有不在 26～29 个循环范围内的 Ct（循环阈）值，从 SDS 软件中打印出样品及其对应 IPC 的扩增图表（ΔRn vs. 循环），并放在合适的案例文件夹中（见疑难解答）。

## 注释

（22）将每个样品的信号强度与从人类 DNA 对照组中产生的标准曲线的信号强度进行比较，将结果除以 2 得到 ng/μL 结果，获得每个样品的 DNA 含量。

（23）在“Results”表中，选择标准曲线设置以绘制标准曲线，从“Detector”下拉列表中选择“Quantifiler Human”。确认标准曲线的斜率为 −2.9 ~ −3.3 并且 $R^2$ 值大于 0.98。

（24）检查并确认对于 DNA 浓度的每减少三倍，标准样的 Ct 值增加约 1.5 个循环。

（25）要检测标准样和观察样品的扩增图，在“Results”列中选择“Amplification Plot”框。从“Detector”下拉菜单中选择“Quantifiler Human”。选择想要观察的孔。

（26）确认 NTC 具有大大高于 38 个循环的 Ct 值或未能超过阈值。

（27）如一个样品显示未测定数值，而且 IPC 超过 26～29 的阈值，说明样品中没有可检测的 DNA。

# 疑难解答

问题［步骤（21）］：IPC 具有超过 26～29 个循环的 Ct 值。

解决方法：

对于所有的标准样、对照样品和检测样品，其 Ct 值应该在 26～29 个循环之间。某一 IPC 具有的 Ct 值在这一范围之外，或有一个没超过基线，表明样品被 PCR 抑制剂或过多的 DNA 所抑制，也有可能同时被两者抑制。

（1）稀释并重新对样品定量。

（2）IPC 值基于 2 μL 提取物。增加模板体积可能增强抑制效应。

（3）如果 IPC 积累的荧光从未超过未使用的探针的荧光（通常出现在 6 个循环左右），而且样品具有比 50 ng 标准样更高的 Ct 值，则 IPC 因为反应体系中有过多的未知 DNA 而失败。IPC 所需的最终浓度为 5 ng，如果未知 DNA 最终浓度高出 20 倍或更多（100 ng 或更多），则 IPC 会因为 IPC 中模板 DNA 的随机扩增而失败。

（4）如果 IPC 积累的荧光达到一定值，然后停留在那里而不超过阈值，同时未知样品积累的荧光超过阈值达 36 个循环，则 IPC 是因为 PCR 抑制剂的存在而失败。

（5）如果 IPC 失败，而未知样品直到 28.3 个循环仍未超过阈值（用步骤（3）中的样品进行完整的 7 次循环），则 IPC 的失败是由于 PCR 抑制剂和高 DNA 浓度的共同作用所致。最可能的是反应体系中有 PCR 抑制剂，但是最终被高浓度的 DNA 模板所打破。然而，这将导致 350pg/μL 的假阳性值。

问题［步骤（23）］：标准曲线并不符合指南的叙述。

解决方法：

（1）通过设定为未被使用的孔，可清除不超过两组的异常数据。欲将标准点设为未被使用，为该孔点击“Well Inspector”并确认“Omit Well”。重新分析数据。

（2）如果清除这些数据点并未改善标准曲线，实验无效。需要重复实验。

# 参考文献


Alarcón M., Cantor R.M., Liu J., Gilliam T.C., Geschwind D.H., Autism Genetic Resource Exchange Consortium. 2002. Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families. *Am. J. Hum. Genet.* **70:** 60–71.

Barker D.L., Hansen M.S., Faruqi A.F., Giannola D., Irsula O.R., Lasken R.S., Latterich M., Makarov V., Oliphant A., Pinter J.H., et al. 2004. Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Res.* **14:** 901–907.

Bergen A.W., Haque K.A., Qi Y., Beerman M.B., Garcia-Closas M., Rothman N., and Chanock S.J. 2005. Comparison of yield and genotyping performance of multiple displacement amplification and OmniPlex whole genome amplified DNA generated from multiple DNA sources. *Hum. Mutat.* **26:** 262–270.

Caputo J.L., Thompson A., McClintock P., and Reid Y.A. 1991. An effective method for establishing human B lymphoblastic cell lines using Epstein-Barr Virus. *J. Tissue Culture Meth.* **13:** 39–44.

Dean F.B., Nelson J.R., Giesler T.L., and Lasken R.S. 2001. Rapid amplification of plasmid and phage dna using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11:** 1095–1099.

Dean F.B., Hosono S., Fang L., Wu X., Faruqi A.F., Bray-Ward P., Sun Z., Zong Q., Du Y., Du J., et al. 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci.* **99:** 5261–5266.

Dunphy C.H. 2006. Gene expression profiling data in lymphoma and leukemia: Review of the literature and extrapolation of pertinent clinical applications. *Arch. Path. Lab. Med.* **130:** 483–520.

Fathallah-Shaykh H.M. 2005. Microarrays: Applications and pitfalls. *Arch. Neurol.* **62:** 16669–16672.

Feigelson H., Rodriguez C., Robertson A., Jacobs E., Calle E., Reid Y., and Thun M.J. 2001. Determinants of DNA yield and quality from buccal cell samples collected with mouthwash. *Cancer Epidemiol. Biomarker Preven.* **10:** 1005–1008.

Garcia M.E., Blanco J.L., Caballero J., and Gargallo-Viola D. 2002. Anticoagulants interfere with PCR used to diagnose invasive aspergillosis. *J. Clin. Microbiol.* **40:** 1567–1568.

Garcia-Closas M., Egan K., Abruzzo J., Newcomb P., Titus-Ernstoff L., Franklin T., Bender P.K., Beck J.C., Le Marchand L., Lum A., et al. 2001. Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash. *Cancer Epidemiol. Biomarker Preven.* **10:** 687–696.

Glasel J.A. 1997. Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios. *Biotechniques* **18:** 62–63.

Gribble S., Ng B.L., Prigmore E., Burford D.C., and Carter N.P. 2004. Chromosome paints from single copies of chromosomes. *Chromosome Res.* **12:** 143–151.

Heath E.M., Morken N.W., Campbell K.A., Tkach D., Boyd E.A., and Strom D.A. 2001. Use of buccal cells collected in mouthwash as a source of DNA for clinical testing. *Arch. Pathol. Lab. Med.* **125:** 127–133.

Holmans P., Zubenko G.S., Crowe R.R., DePaulo Jr., J.R., Scheftner W.A., Weissman M.M. Zubenko W.N., Boutelle S., Murphy-Eberenz K., MacKinnon D., et al. 2004. Genomewide significant linkage to recurrent, early-onset major depressive disorder on chromosome 15q. *Am. J. Hum. Genet.* 74: 1154–1167.

Hosono S., Faruqi A.F., Dean F.B., Du Y., Sun Z., Wu X., Du J., Kingsmore S.F. Egholm M., and Lasken R.S. 2003. Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* **5:** 954–964.

King I.B., Satia-Abouta J., Thornquist M.D., Bigler J., Patterson R.E., Kristal A.R., Shattuck A.L., Potter J.D., and White E.

2002. Buccal cell DNA yield, quality, and collection costs: Comparison of methods for large-scale studies. *Cancer Epidemiol. Biomark. Prev.* **11:** 1130–1133.

Lahiri D.K., Bye S., Nurnberger J.I., Jr., Hodes M.E., and Crisp M. 1992. A non-organic and non-enzymatic extraction method gives higher yields of genomic DNA from whole blood samples than do nine other methods. *J. Biochem. Biophys. Methods* **25:** 193–205.

Langmore J.P. 2002. Rubicon Genomics, Inc. *Pharmacogenomics* **3:** 557–560.

Lee C.I.P., Leong S.H., Png A.E.H., Choo K.W., Syn C., Lim D.T.H., Law H.Y., and Kon O.L. 2006. An isothermal method for whole genome amplification of fresh and degraded DNA for comparative genomic hybridization, genotyping and mutation detection. *DNA Res.* **13:** 77–88.

Li J., Harris L., Mamon H., Kulke M.H., Liu W.H., Zhu P., and Makrigiorgos G.M. 2006. Whole genome amplification of plasma-circulating DNA enables expanded screening for allelic imbalance in plasma. *J. Mol. Diagn.* **8:** 22–30.

Little S.E., Vuononvirta R., Reis-Filho J.S., Natrajan R., Iravani M., Fenwick K., Mackay A., Ashworth A., Pritchard-Jones K., and Jones C. 2005. Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA. *Genomics* **87:** 298–306.

Lovmar L., Fredriksson M., Liljedahl U., Sigurdsson S., and Syvanen A.C. 2003. Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. *Nucleic Acids Res.* **31:** e129.

Lum A. and Le Marchand L. 1998. A simple mouthwash method for obtaining genomic DNA in molecular epidemiological studies. *Cancer Epidemiol. Biomark. Prev.* **7:** 719–724.

Luthra R. and Medeiros L.J. 2004. Isothermal multiple displacement amplification: A highly reliable approach for generating unlimited high molecular weight genomic DNA from clinical specimens. *J. Mol. Diagn.* **6:** 236–242.

Madisen L., Hoar D.I., Holroyd C.D., Crisp M., and Hodes M.E. 1987. DNA banking: The effects of storage of blood and isolated DNA on the integrity of DNA. *Am J. Med. Genet.* **27:** 379–390.

Manchester K.L. 1995. Value of A260/280 ratio for measurement of purity of nucleic acids. *Biotechniques* **19:** 208–210.

———. 1996. Use of UV methods for measurement of protein and nucleic acid concentrations. *Biotechniques* **20:** 968–970.

McEwen J.E. and Reilly P.R. 1994. Stored Guthrie cards as DNA banks. *Am. J. Hum. Genet.* **55:** 196–200.

Neitze H. 1986. A routine method for the establishment of permanent growing lymphoblastoid cell lines. *Hum. Genet.* **73:** 320–326.

Nicklas J.A. and Buel E. 2003. Development of an *Alu*-based real-time PCR method for quantitation of human DNA in forensic samples. *J. Forensic Sci.* **48:** 936–944.

Paez J.G., Lin M., Beroukhim R., Lee J.C., Zhao X., Richter D.J., Gabriel S., Herman P., Sasaki H., Altshuler D., et al. 2004. Genome coverage and sequence fidelity of ϕ29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32:** e71.

Pask R., Rance H.E., Barratt B.J., Nutland S., Smyth D.J., Sebastian M., Twells R.C., Smith A., Lam A.C., Smink L.J., et al. 2004. Investigating the utility of combining Φ29 whole genome amplification and highly multiplexed single nucleotide polymorphism BeadArray™ genotyping. *BMC Biotechnol.* **4:** 15.

Prasad H.C., Zhu C.B., McCauley J.L., Samuvel D.J., Ramamoorthy S., Shelton R.C., Hewlett W.A., Sutcliffe J.S., and Blakely R.D. 2005. Human serotonin transporter variants display altered sensitivity to protein kinase G and p38 mitogen-activated protein kinase. *Proc. Natl. Acad. Sci.* **102:** 11545–11550.

Quackenbush J. 2006. Microarray analysis and tumor classification. *N. Engl. J. Med.* **354:** 2463–2472.

Quantifiler Human DNA and Y Human Male Quantification Kit User's Manual. 2004. Applied Biosystems, Foster City, California.

Reis-Filho J.S., Westbury C., and Pierga J.Y. 2006. The impact of expression profiling on prognostic and predictive testing in breast cancer. *J. Clin. Pathol.* **59:** 2225–2231.

Reyes F., Gourdin M.F., Lejonc J.L., Cartron J.P., Gorius J.B., and Dreyfus B. 1976. The heterogeneity of erythrocyte antigen distribution in human normal phenotypes: An immunoelectron microscopy study. *Br. J. Haematol.* **34:** 613–621.

Rylander-Rudqvist T., Hakansson N., Tybring G., and Wolk A. 2006. Quality and quantity of saliva DNA obtained from the self-administrated oragene method—A pilot study on the cohort of Swedish men. *Cancer Epidemiol. Biomark. Prev.* **15:** 1742–1745.

Steinberg K., Beck J., Nickerson D., Garcia-Closas M., Gallagher M., Caggana M., Reid Y., Cosentino M., Ji J., Johnson D., et al. 2002. DNA banking for epidemiological studies: A review of current practices. *Epidemiology* **13:** 246–254.

Suarez B.K., Duan J., Sanders A.R., Hinrichs A.L., Jin C.H., Hou C., Buccola N.G., Hale N., Weilbaecher A.N., Nertney D.A., et al. 2006. Genomewide linkage scan of 409 European-ancestry and African American families with schizophrenia: Suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample. *Am. J. Hum. Genet.* **78:** 315–333.

Telenius H., Carter N.P., Bebb C.E., Nordenskjold M., Ponder B.A., and Tunnacliffe A. 1992. Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13:** 718–725.

Thorstenson Y.R., Hunicke-Smith S.P., Oefner P.J., and Davis R.W. 1998. An automated hydrodynamic process for controlled, unbiased DNA shearing. *Genome Res.* **8:** 848–855.

Tomlins S.A., Mehra R., Rhodes D.R., Shah K.B., Rubin M.A., Bruening E., Makarov V., and Chinnaiyan A.M. 2006. Whole transcriptome amplification for gene expression profiling and development of molecular archives. *Neoplasia* **8:** 153–162.

Willinger W.W., Mackey M., and Chomczynski P. 1997. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *Biotechniques* **22:** 474–481.

Zhang L., Cui X., Schmitt K., Hubert R., Navidi W. and Arnheim N. 1992. Whole genome amplification from a single cell: Implications for genetic analysis. *Proc. Natl. Acad. Sci.* **89:** 5847–5848.

# 10　中通量实验室规模的基因分型方法

Stuart J. Macdonald

*Department of Ecology and Evolutionary Biology and Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66045*



## 简介

近几年，实验技术迅猛发展，超高通量基因分型平台的开发利用使我们可以同时对数以千计的 SNP 位点进行基因分型（见综述，第 13 章）。这些技术给人类医学遗传学家带来福音，先前难以想象的实验逐渐成为常规工作。然而，很多研究不需要检测数千个 SNP，因而不适合使用超高通量的基因分型平台。这篇综述旨在为致力于基因分型数千至数十万个 SNP 的研究提供指导。

本章首先概述各种已有的基因分型平台，着重叙述各检测方法鉴别等位基因的原理以及等位基因特异性基因分型产物的检测方法。就每一种方法，将指出该方法在操作时的多重兼容性，即单个反应能对多少个 SNP 基因分型。这里介绍的基因分型方法大多数都以涵盖靶 SNP 位点的 PCR 扩增产物为模板。PCR 的多重性即一个 PCR 反应扩增出多个独立片段，可以提高基因分型的通量。多重 PCR 的设计和操作不算繁琐但也不

常用，本综述不专就此方法做过多讨论。第二节制定了基因分型平台的选择标准，并从理论上提出 4 种方案（每种方案要求不同数量的靶 SNP 和个体样本）。最后，将讨论靶 SNP 以外的其他 DNA 多态性对基因分型准确性的影响。下一章介绍一系列详细的基因分型操作规程进行介绍。

## 基因分型检测方法

### 测序

对于某些课题，没有必要花费时间和精力去探索基因分型技术。例如，在待分析样本和位点数都很少的情况下，对所有样本的目的片断测序可能更简便。测序可以得到该区段的所有多态性信息，不仅是最初选择的几个可预见的 SNP。此外，通过测序还能发现一个 SNP 位点罕见的第三种（甚至第四种）等位基因。相反，下面将要讨论的许多其他基因分型方法都是建立在所有目标 SNP 均为变等位基因的前提下。

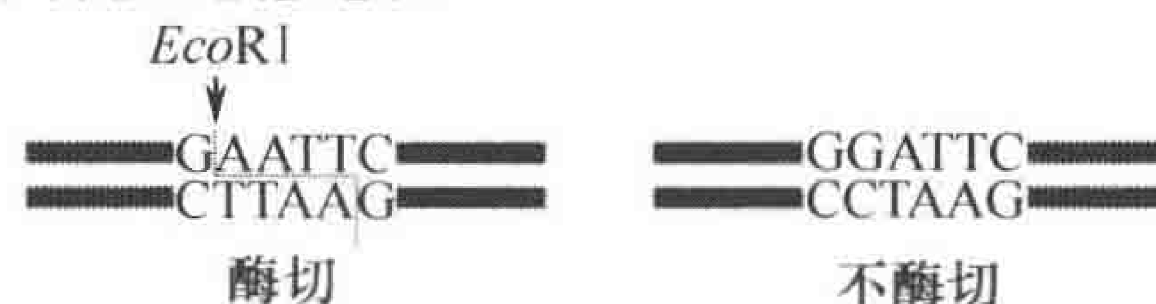
### 等位基因特异性 PCR

通过等位基因特异性 PCR 可以成功地对 SNP 位点进行基因分型（Newton et al. 1989；Li et al. 1990；Bottema and Sommer 1993），有时称为扩增阻滞突变系统（amplification refractory mutation system，ARMS；Newton et al. 1989）。常规 PCR 扩增一段 DNA，所用的上、下游引物与靶序列完全配对，然而等位基因特异性 PCR（图 10-1A）采用等位基因特异的两条上游引物，两者 3′端核苷酸不同，一个对野生型等位基因特异，另一个对突变型等位基因特异。在适当的严格杂交条件下，与模板不完全匹

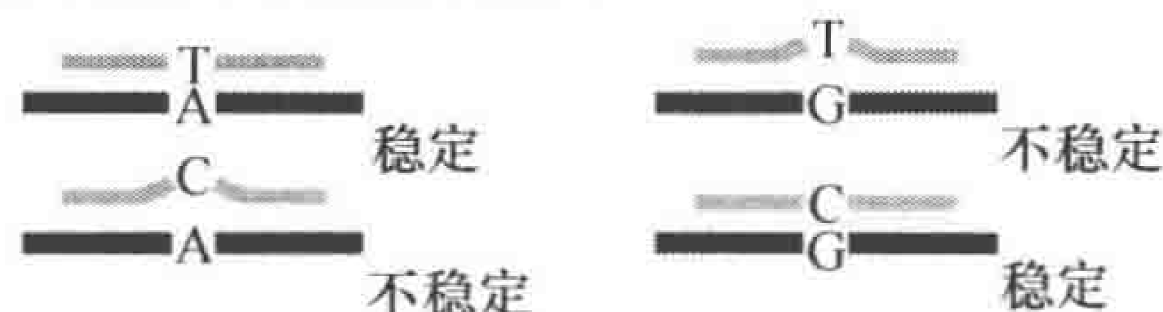
A 等位基因特异性PCR



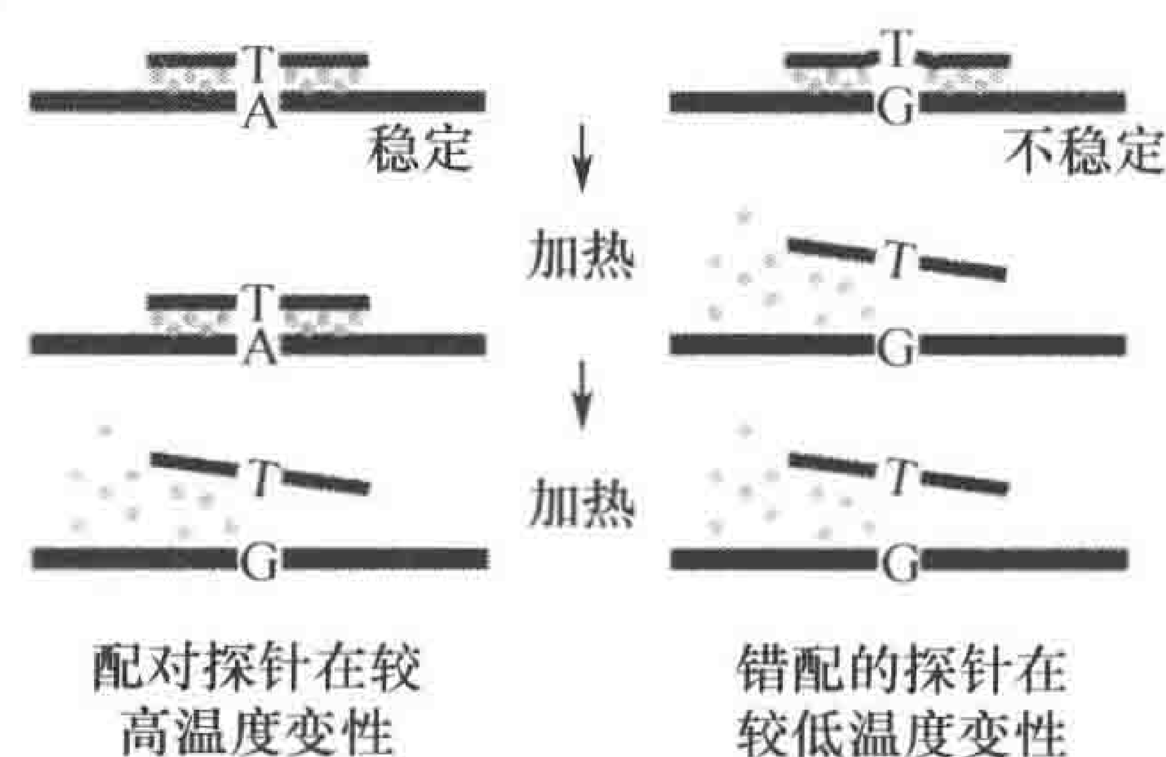

B 限制性片段长度多态性



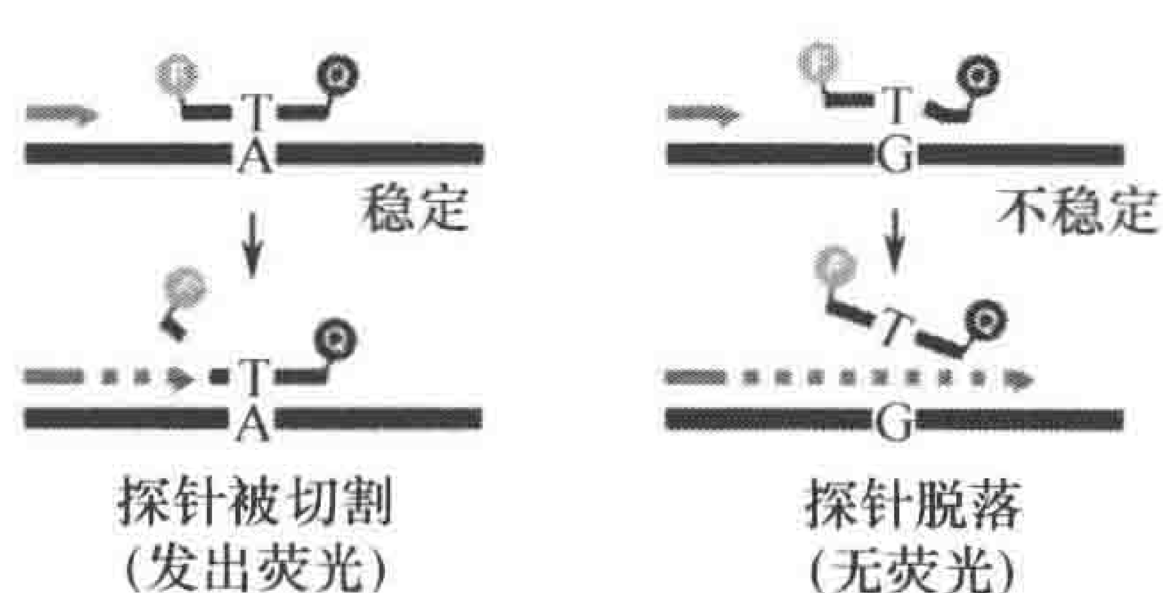

C 等位基因特异性寡核苷酸杂交




D 动态等位基因特异性杂交




E TaqMan



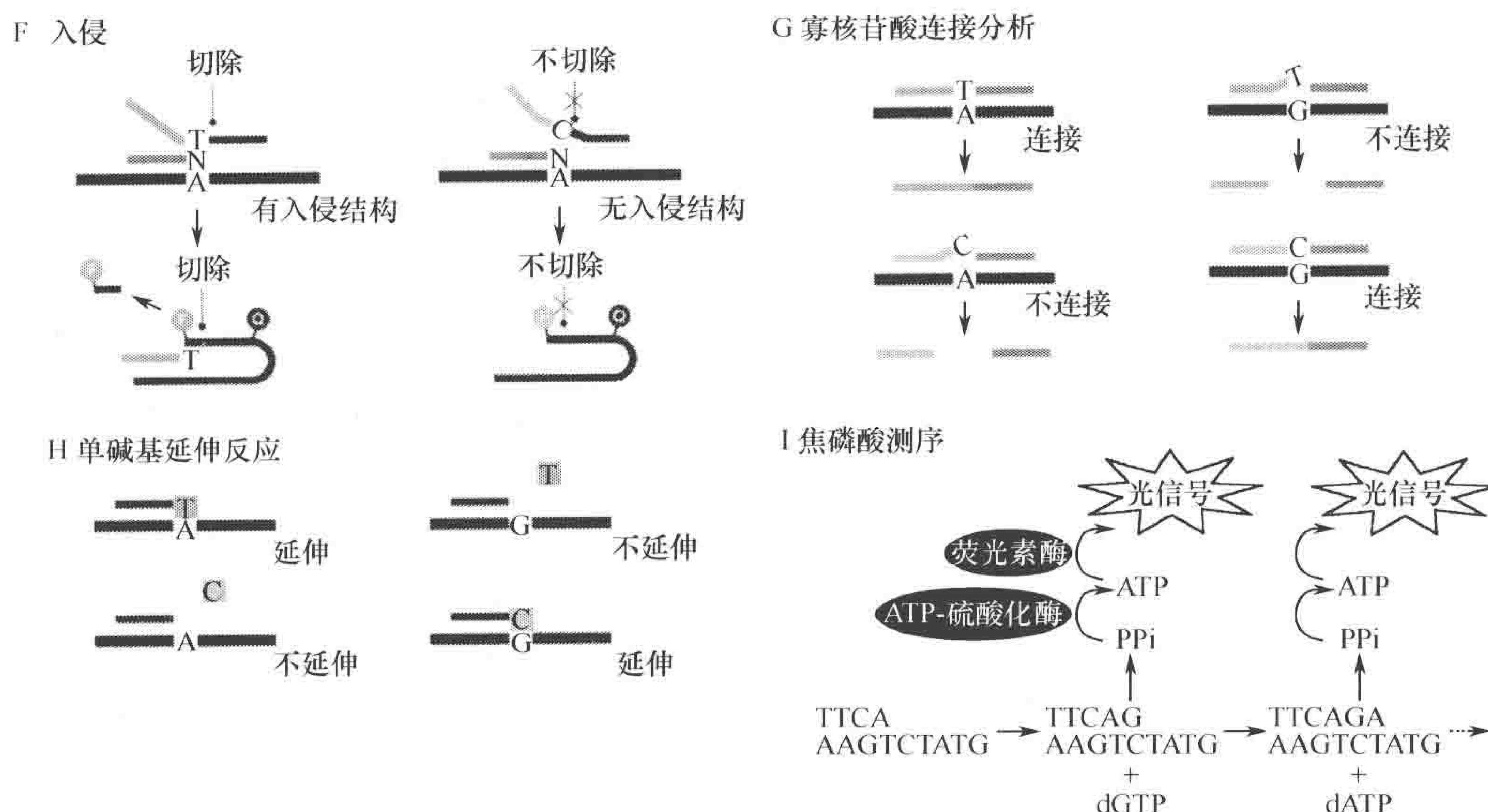



图 10-1 基因分型方法概述。A. 等位基因特异性 PCR 如果等位基因特异性寡核苷酸（红色/绿色）与靶序列（黑色）互补，寡核苷酸将被延伸；否则，没有 PCR 产物。B. 限制性片段长度多态性 如果 SNP 等位基因保留限制性位点，限制性内切酶（如 *Eco*RⅠ）特异性切割 PCR 产物片断。C. 等位基因特异性寡核苷酸杂交 标记的等位基因特异性寡核苷酸（红色/绿色）与固定在固相支持物表面的靶 DNA（黑色）杂交。只有完全匹配，ASO-模板复合体才能稳定结合。D. 动态等位基因特异性杂交 在双链 DNA 特异性染料存在的情况下，一条等位基因特异性引物与靶序列杂交。然后加热探针-靶序列复合物，与完全匹配探针相比，在较低温度下错配探针就变性，失去荧光。E. TaqMan TaqMan FRET 探针 5′端连有荧光基团，3′端连有荧光淬灭基因（图 10-2），如果探针与靶序列完全匹配，则杂交形成稳定的复合体。图中只显示了两条探针中的一条。如果完全匹配，引物（蓝色）延伸，探针被 Taq DNA 聚合酶降解，释放荧光信号。如果探针与靶序列错配，探针-模板复合物不稳定，探针被置换，并保持完整，没有荧光信号发出。F. Invader 分析法 Invader 寡核苷酸（蓝色）和一对等位基因特异性寡核苷酸与靶序列杂交。图中只显示了 SNP 两个等位基因中的一个。当等位基因特异性寡核苷酸与模板配对，形成特异的 Invader 结构，等位基因特异性寡核苷酸释放出 5′端尾序列（红色）。这段尾序列可以作为 Invader 寡核苷酸参与 FRET 盒的二级 Invasive 剪切反应，当 FRET 盒末端被切除，释放荧光信号。G. 寡核苷酸连接分析 一对等位基因特异性寡核苷酸（红色/绿色）及一条位点特异性寡核苷酸（蓝色）能够与靶序列杂交。两种完全匹配的寡核苷酸紧密并行与模板形成杂交体，两探针共价连接。很多种方法都能区分连接产物和未连接寡核苷酸。H. 单碱基延伸 一条基因分型探针与靶 SNP 上游退火，3′端结束在 SNP 位点前一碱基处，DNA 聚合酶募集一个与 SNP 等位基因互补的标记核苷酸，探针只延伸一个碱基。I. 焦磷酸测序 基因分型探针（粗体）与靶 SNP（s）上游退火，依次加入 4 种脱氧核糖核苷酸（dATP，dCTP ，dGTP，dTTP），如果掺入互补的 dNTP，发生一系列酶催化反应，引起光信号的释放

配的上游引物将不能退火，不能生成 PCR 产物。每个样本，都用两个等位基因特异的上游引物分别与共同的下游引物进行 PCR 扩增，通过标准琼脂糖凝胶电泳，溴化乙锭染色观察 PCR 产物的有无，即可确定该 SNP 位点的基因型。若改变其中一条等位基因上游引物的长度，将三条引物加到同一个 PCR 反应管里，通过 PCR 产物的长度即可确定该 SNP 位点的基因型（Li et al. 1990；Bottema and Sommer 1993）。四引物 ARMS 法（Ye et al. 2001），同时使用 4 条 PCR 引物：两条等位基因 3′端碱基特异性的内引物，分别与 DNA 模板的两条链互补；一对完全匹配的外引物，两条外引物与目标 SNP 距离不同。所有样本经 PCR 扩增都将得到由两条外引物扩增而来的产物（作为 PCR 反应的阳性对照），另外包括由各样本 SNP 位点基因型决定的一条或两条带。

熔解温度曲线法（Germer and Higuchi 1999；Wang et al. 2005）是建立在等位基因特异性 PCR 基础上的一种基因分型方法，区别在于该方法是利用等位基因特异性 PCR 产物的熔解温度差异进行基因分型，而不是通过电泳检测。共同引物位于 SNP 位点下游 20bp 左右，使目的扩增产物片段尽量短。为了在两个较短等位基因特异性 PCR 产物间产生可检测的温度差异，可以在两条等位基因特异性引物末端掺入一段不同长度（6nt：14nt）富含 GC 的序列。该 PCR 体系包含三条引物以及双链特异性 DNA 染料 SYBR Green Ⅰ。PCR 结束后，通过实时定量 PCR 仪监测随 PCR 产物解链产生的荧光信号下降值，绘制熔解曲线。由于各自 GC 尾巴长短不同，两种等位基因特异性 PCR 产物产生的熔解曲线不同因而得以区分。

尽管对于部分 SNPs 位点来说，等位基因特异性 PCR 是一种经济有效的基因分型手段，但是它不能很好的应用于大量 SNP 的分型项目。首先，它不能在单个实验中同时完成多个反应。其次，它要求严格的 PCR 反应条件，以防止错配引物与完全匹配引物等效率退火而产生假阳性。一套 PCR 反应条件不能广泛应用于所有检测。

**限制性片段长度多态性**（restriction fragment length polymorphism，RFLP）

这是一种最简单最直接的 SNP 等位基因检测方法，利用限制性内切核酸酶能识别特异性 DNA 序列的特点实现 SNP 位点的等位基因识别（Saiki et al. 1985）。PCR 扩增包含靶 SNP 在内的一段基因组 DNA，然后用限制性内切核酸酶酶切 PCR 产物，只有特异识别位点能被切割（图 10-1B）。该方法能对已知 SNP 进行有效的分析，但不是所有的 SNP 都能找到合适的限制性内切核酸酶。在设计 PCR-RFLP 基因分型实验时可以利用一些辅助软件（SNP cutter，http://bioinfo. bsd. uchicago. edu/SNP_cutter. htm；Zhang et al. 2005）。酶切产物在溴化乙锭-琼脂糖凝胶上电泳，根据片段长度的变化即可推知基因型，操作简便、经济有效。或者，可以将酶切产物作荧光标记，然后用毛细管测序仪进行分辨。可以将初始 PCR 引物标记，或者可以将酶切产物荧光标记（Lazzaro et al. 2002）。荧光 PCR-RFLP 检测能达到中等程度的多重性。多重 PCR，每种产物用不同的荧光基团标记，多重限制性酶切，产生大小不同、荧光颜色不同的各种特异性片段（Thomas et al. 1999）。

PCR-PFLP 实验如果没有得到预计的酶切产物，可能是因为该位点为另外一个等位基因，也可能是实验操作失误造成的。因此，有必要用已知基因型的对照样本做

PCR-RFLP 检测失败率，同时对部分或所有样本重复检测。此方法的局限性是 PCR-RFLP 检测不适用于大规模基因分型。

## 等位基因特异性杂交

有几种基因分型技术都是基于等位基因特异性寡核苷酸（allele-specific oligonucleotide，ASO）探针杂交（图 10-1C）的原理。由于没有酶的参与，杂交是最简单的 SNP 基因分型策略之一。较短的寡核苷酸探针与被检测突变所在区域的 DNA 序列互补，在高度严格的杂交和洗脱条件下，只有完全匹配才能稳定杂交，只要有一个碱基不匹配，就不能形成稳定的杂交体（Wallace et al. 1979；Conner et al. 1983）。该方法最简单的应用是 PCR 扩增靶基因组 DNA 区段，然后将 PCR 产物转移到尼龙膜上，再与放射标记的 ASO 探针杂交（Saiki et al. 1986）。ASO 探针通常比较短（10～20 个寡核苷酸），除了中央一个用于 SNP 鉴别的等位基因特异性核苷酸之外，其他碱基组成相同；当探针用于插入或缺失多态性分析时，探针中间包含一串特异的核苷酸。因为在同一张膜上可以阵列大量样本，所以 ASO 是一种处理大样本的有效方法。关键的技术难点是如何优化适合每个 SNP 的严格杂交和洗脱条件，以保证探针杂交的特异性。这一难点潜在地限制了这种方法的适用范围，有些 SNP 就不能用这种简单的方法检测了。

动态等位基因特异性杂交（dynamic allele-specific hybridization，DASH）是通过直接监测荧光标记探针与目标序列形成的杂交体变性特征而实现 SNP 基因分型的技术（Howell et al. 1999；Prince et al. 2001）。双链 DNA 特异性染料（如溴化乙锭、SYBR GreenⅠ）嵌入 DNA 双链中，随着温度升高，监测染料荧光强度变化绘制出 DNA 熔解曲线（Ririe et al. 1997）。当温度接近熔解温度时，荧光强度迅速减弱。DASH 检测就是利用对变性温度变化过程的观察（图 10-1D）。用生物素标记一条引物，PCR 扩增的目的片段，PCR 产物经生物素结合于链亲和素包被的固体反应表面（如微量反应板的反应孔）。加入探针和荧光染料，探针长 15～21 个核苷酸，正中一个碱基与 SNP 的一种等位基因完全匹配。探针和结合于固体反应表面的单链 PCR 产物在低温条件下杂交，荧光染料可特异性地插入 DNA 双链中。发射的荧光强度与杂交体的形成量相关。在实时 PCR 仪中持续缓慢升温，同时监测荧光强度，即能反映杂交体的熔解温度。探针与靶序列之间有一个碱基错配就能明显降低熔解温度，因此，SNP 位点的基因型很容易鉴别（Howell et al. 1999）。与基于阵列的 ASO 法类似，探针设计是 DASH 检测成功与否的一个决定因素。2001 年，Prince 等建立了探针设计标准。因为一个反应仅能检测一个等位基因特异性探针，所以分析一个 SNP 位点需要两个反应，由于 DASH 法的多重性低，因而其只能用于分析少量 SNP。为了突破这个局限，建立了基于阵列的多重 DASH 基因分型法，称为 DASH-2（Jobs et al. 2003）。将单链等位基因特异性 PCR 产物固定在芯片上，而不再使用微量反应板。

5′核酸酶法，又称 TaqMan 探针法，是另一种应用更为广泛的基于等位基因特异性杂交原理设计的 SNP 检测技术（图10-1E）（Livak 1999；De La Vega et al. 2005）。因为 PCR 是该方法的核心部分，TaqMan 法能直接以基因组 DNA 为模板，对 SNP 和短的插入/缺失多态性位点进行基因分型。但实现其多重性也很难做到。每检测一个 SNP 位

点需要一对引物扩增靶 SNP 所在序列，还需要一对等位基因特异性荧光共振能量转移（fluorescence resonance energy transfer，FRET）探针，其 5′端分别连有两种不同的荧光染料，3′端连有通用的荧光淬灭基因（图 10-2）。一个完整的探针，淬灭基因和荧光基团在空间上特别靠近而产生荧光淬灭（Livak et al. 1995）。PCR 体系包括引物、标记探针、基因组 DNA 和具有 5′→3′核酸酶活性的 Taq DNA 聚合酶等，在 PCR 仪上进行热循环反应。在 PCR 扩增过程中，TaqMan 探针与相应的靶序列配对，Taq DNA 聚合酶将引物延伸，当 Taq DNA 聚合酶与杂交的探针相遇时，发挥 5′→3′核酸酶活性，切割探针，使荧光基因与淬灭基因分离而发出荧光。错配的探针不能与靶序列稳定配对，就不会被 Taq DNA 聚合酶切割，而且促进了探针置换。与完全匹配的探针相比，减少了探针被切割的时间。随着 PCR 反应的进行，荧光信号逐渐增强，根据样本的基因型不同，可能产生一种或两种荧光信号，用实时 PCR 仪即可检测。与等位基因特异性杂交技术相似，TaqMan 探针法区分特异性等位基因成败的关键也在于实验设计，表 10-1 列出了探针和引物的设计原则（源自 Livak 1999），另外，网上提供用户实验设计服务（http://www.appliedbiosystems.com）。

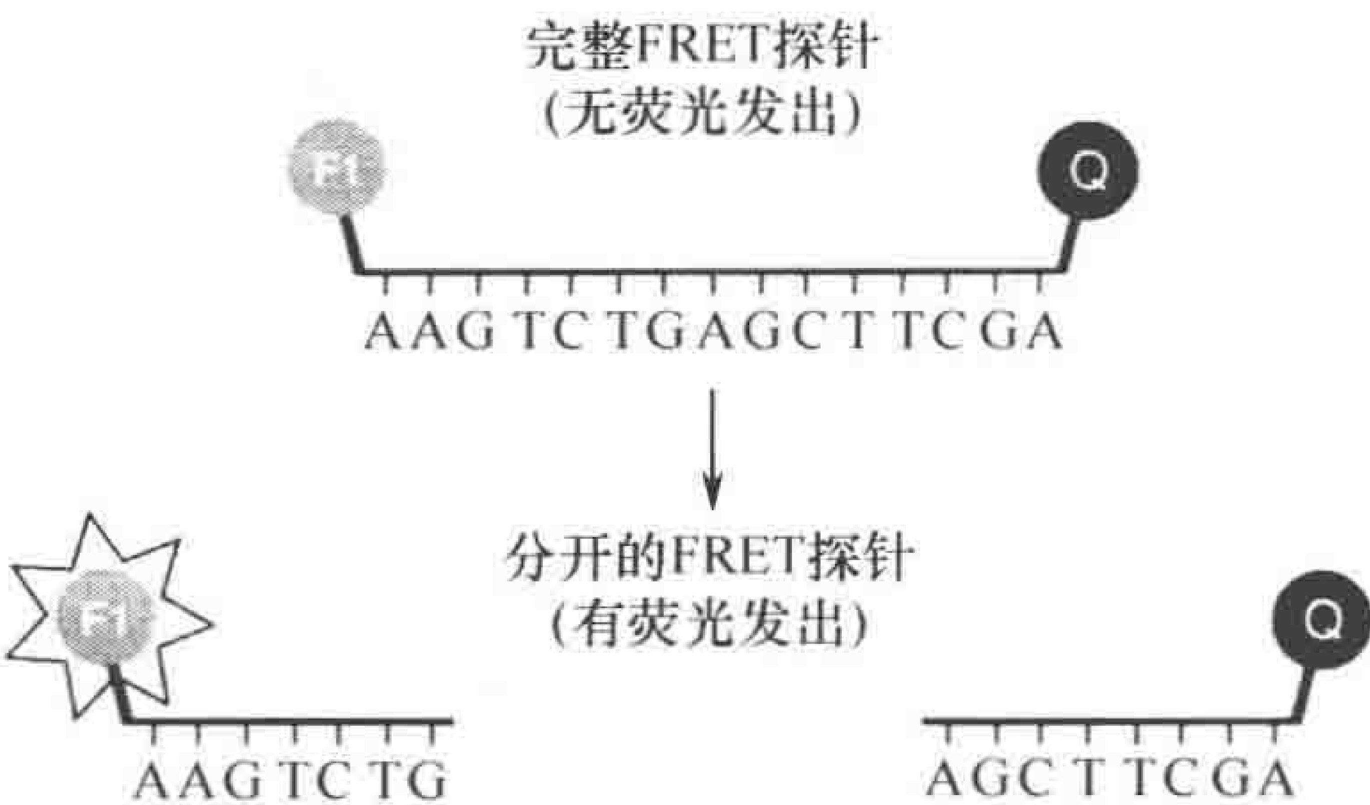



图 10-2　FRET 探针 FRET 探针包含一段与特异性 DNA 序列互补的短核苷酸序列，通常围绕中间一个特异性 SNP 等位基因。FRET 探针 5′端标记一个荧光基团，3′端标记淬灭基团。只有荧光剂的发射光谱与淬灭剂的激发光谱重叠时，荧光基因吸收的激发能不以荧光形式释放，而是通过共振传递给淬灭剂。荧光信号淬灭的效力取决于荧光基团与淬灭剂之间的距离：假如淬灭基团与荧光基团相互靠近，探针无荧光发出。然而当两基团分离后（如因为探针被剪切而分离），荧光基团发出荧光

**表 10-1　TaqMan 法实验设计原则**

| 引物 | 探针 |
|---|---|
| GC 含量 30%～80% | |
| 无 3 个以上连续的 G | |
| $T_m$：58～60℃ | $T_m$：65～67℃[a] |
| 上、下游引物与交叠的探针应该尽可能靠近 | |
| 3′端最后 5 个核苷酸应只包含 1 或 2 个 G 或 C | 选择 C 比 G 含量多的链做探针 |

注：经允许，表格引自 Livak[1999（Elsevier）]，并经修改。

a 熔解温度（$T_m$）由 Primer Express 软件提供（Applied Biosystems）。

## Invasive 寡核苷酸切割（Invasive oligonucleotide cleavage）

Invader 分析使用一个耐热的、结构特异的内切核酸酶，根据结构而不是序列在特殊位点上切割核酸分子，是一种等温非“PCR”的 DNA 和 RNA 定性和定量检测方法。可简单、快速、准确检测基因突变和 SNP。Invader 分析法是通过一对寡核苷酸与目标序列杂交形成的特异性 3 股螺旋结构被切割酶识别并剪切而实现基因分型的，与核酸序列无关（图 10-1F）（Olivier 2005）。每一反应体系包括两种寡核苷酸：一种是 Invasive 寡核苷酸，位于待测 SNP 位点上游，与目标序列互补，3′端的碱基与 SNP 位点重叠（尽管 3′端碱基与酶的切割作用无关）；另一种是等位基因特异性寡核苷酸，包含与 SNP 位点等位基因互补的核苷酸，从 SNP 位点沿目标序列向下游延伸，另外，含有一段 5′端尾序列（所谓的“瓣”）与目标序列不互补。两种寡核苷酸同时与目标序列杂交形成特异性 3 股螺旋结构，在 SNP 位点处重叠，从古细菌属分离的热稳定瓣内切酶（flap endonu cleases，FEN）能识别这种特殊结构并将其切除，释放出等位基因特异性寡核苷酸 5′端尾序列（Lyamichev et al. 1999）。如果等位基因特异性寡核苷酸在 SNP 位点处存在错配，就不会形成重叠结构，也就不会发生剪切反应。等位基因特异性寡核苷酸可以是 FRET 探针（图 10-2），5′端分别连有两种不同的荧光染料。被切除的 5′端尾序列带着等位基因特异性的荧光标记，当它与共存的淬灭剂分离后就能释放荧光而被检测。

Invader 分析是等温反应，没有热循环的过程。为了保证等位基因特异性荧光达到可检测水平，选择一个合适的反应温度很重要。调整两个等位基因特异性寡核苷酸的熔解温度，使其在反应温度下，与目标序列退火和分离持续交替进行。相反，Invasive 寡核苷酸则要设计成与目标序列始终退火。这样，才能连续产生 Invader 结构，连续发生切割反应，荧光释放量才会随着反应进行而不断增强。

上述这种模式的 Invader 分析法有一个缺点，即每分析一个 SNP 位点，都需要两个很贵的 FRET 探针。系列 Invasive 信号扩增反应（the serial invasive signal amplification reaction，SISAR）（Hall et al. 2000）解决了这个问题。在该反应中，一对等位基因特异性寡核苷酸，与目标序列互补的部分不变，5′端尾序列碱基组成不同。这对尾序列是通用的，对另外一对通用的 FRET 盒特异（图 10-1F）。Invader 分析反应过程中被切除释放的尾序列可以作为 Invader 寡核苷酸与 FRET 盒进行二级反应。FRET 盒末端和被切除尾序列形成的 Invader 寡核苷酸结合，形成特异的 Invader 结构，使 FRET 盒末端被切除，荧光基团与共存的淬灭基团分离，释放荧光信号。这两种模式的 Invader 分析法均可采用 PCR 扩增产物作为模板，但是 SISAR 直接以基因组 DNA 为模板分析更加灵敏（Hall et al. 2000），尽管这需要大量的基因组 DNA（20～100ng；Olivier 2005）。这两种分析模式都不能实现较高的多重性，因而不能在单个反应中完成多个 SNP 的分析，但采用固相检测策略可以达到高通量分析。例如，可以将标记的初始 Invasive 探针吸附到微珠上，剪切反应完成后，用流式细胞仪检测微珠上的荧光信号（Rao et al. 2003）。或者将 Invasive 探针固定在基因芯片表面，剪切反应后，用普通成像系统即可读取芯片上的结果（Lu et al. 2002）。

## 寡核苷酸连接分析

寡核苷酸连接分析（oligonucleotide ligation assay，OLA）（Landegren et al. 1998）是利用三条寡核苷酸及一种热稳定 Taq DNA 连接酶进行 SNP 鉴别的技术（图 10-1G）。一对等位基因特异性寡核苷酸互补于待测 SNP 位点一侧序列，根据 3′末位碱基不同鉴别等位基因。一条位点特异性（5′端磷酸化）公共寡核苷酸链互补于待测 SNP 位点另一侧序列，一端紧邻 SNP 位点。与目标 DNA 单链完全互补的等位基因特异性寡核苷酸和公共寡核苷酸紧密并行，与目标 DNA 单链形成杂交体，在 DNA 连接酶的作用下共价连接。OLA 区分完全匹配和错配等位基因特异性寡核苷酸非常有效（Landegren et al. 1998；Luo et al. 1996；Schouten et al. 2002）。OLA 反应在 PCR 仪上进行，循环进行变性和退火（连接）两步反应，连接反应产物逐渐增加达到可检测水平。在一个反应体系中加入多组 OLA 寡核苷酸，使达到相对高的多重性检测水平，可同时检测几十至几百个 SNP 位点或者插入（或缺失）多态性位点。如此高的多重性检测，仅需很少的条件优化工作。

基于 OLA 原理建立了很多基因分型平台，各平台区别在于检测和鉴别等位基因特异性连接产物的方法不同。最简单的一种方法是通过延长其中一条等位基因特异性寡核苷酸链的 5′端序列长度，来改变不同连接产物的电泳迁移率（Barany 1991），连接产物经凝胶电泳可区分，即确定基因型。通过改变连接产物的迁移能力、荧光标记连接反应并利用毛细管测序仪（Grossman et al. 1994；Day et al. 1995；Eggerding 1995；Schouten et al. 2002），可以进一步提高多重性检测能力和样本通量。SNPWave 技术是利用电泳区分片断大小不等的等位基因特异性连接产物的高通量基因分型技术（van Eijk et al. 2004）。它结合了一些扩增片断长度多态性（amplified fragment length polymorphism，AFLP）分析的原理（Vos et al. 1995）。SNPWave 技术用一对较长的（80～130 nt）等位基因特异性开环探针（Nilsson et al. 1994），取代三条基因分型寡核苷酸。开环探针的 5′端和 3′端各有一段序列与目标序列互补，两端之间是一段可变长度的连接序列，5′端紧邻 SNP 位点，3′端核苷酸用于鉴别等位基因。在连接酶的作用下，如果 3′端与靶等位基因完全互补，则 5′端和 3′端发生连接反应，形成闭环 DNA 序列。经过多重 OLA，荧光标记通用 PCR 引物，以闭环探针为模板，扩增出各种长度的等位基因特异性产物，PCR 产物可以通过毛细管测序仪鉴别。SNPWave 技术的优点在于相对较高的多重性检测能力（10～100plex），并且可以直接用基因组 DNA 作为连接反应的模板。但是，长开环探针必须用高压液相色谱（high-pressure liquid chromatography，HPLC）纯化，这就明显提高了每次检测的成本。

用酶联免疫分析（enzyme-linked immunosorbent assays，ELISA）检测 OLA 产物，可以替代电泳，操作简单（Nickerson et al. 1990；Tobe et al. 1996）。该方法将一条等位基因特异性寡核苷酸 5′端标记荧光素，另一条等位基因特异性寡核苷酸 5′端标记地高辛，位点特异性公共寡核苷酸 3′端生物素化。OLA 完成后，连接反应产物将结合到包被有亲和素的微量反应板的反应孔里，然后，分别加入碱性磷酸酶标记的抗荧光素抗体和辣根过氧化物酶标记的抗地高辛抗体。再先后加入碱性磷酸酶底物和辣根过氧化

物酶底物显色，分别用于检测荧光素报告信号和地高辛报告信号。最后用酶标仪进行分光光度法检测，即可判断各样本的基因型。用微量反应板，OLA-ELISA 法可以检测较大的样本量，非常经济有效。然而，这种方法不能实现多重检测，而且，由于要进行几项生化反应，本方法在许多实验室的有效应用受到限制。

通过片段长度及荧光标记鉴别等位基因特异性连接产物限制了多重复用水平，缩小了可用的等位基因检测方法范围。在等位基因特异性基因分型寡核苷酸 5′端增加一段特异性寡核苷酸序列或 DNA 条码是一个有效的解决方案。OLA 反应完成后，每种连接反应产物通过特异的条形码与一个特异性的 SNP 等位基因相对应。寡核苷酸尾端的条形码对连接反应没有干扰，又是通用的（如同样一对条形码可以用于不同复用体系中的不同 SNP 位点检测），所以广泛应用于以 OLA 为基础的各种基因分型技术中。

在等位基因特异性分型寡核苷酸 5′端增加条形码使在阵列上进行 SNP 等位基因检测成为可能。与基于凝胶或毛细管的检测方法相比，阵列方法最大的优点体现在相对容易实现自动化提取数据。Macdonald 等（2005b）设计了一种基于阵列的 OLA 法，一个反应能够检测多达 16 个 SNP 位点。32 条等位基因特异性寡核苷酸，每条寡核苷酸的序列特异性区域上游都包括一段特异的条形码。在等位基因特异性寡核苷酸 5′端及位点特异性寡核苷酸 3′端各有一段标签序列，互补于通用的 PCR 扩增引物。以 PCR 扩增产物为模板进行 OLA 反应，对连接反应产物进行 PCR 扩增，然后将扩增产物转印到尼龙膜上，再加入同位素标记的探针（互补于条形码序列），使其与膜上靶序列杂交。将尼龙膜贴于储存磷光质屏上曝光，即可读取各 SNP 位点的基因型。该方法特别适合大样本检测，因为每张膜上可转印数千个受检者的样本。反向系统是指先在固相支持物上阵列条形码，然后与标记的连接反应产物杂交（Gerry et al. 1999；Banér et al. 2003），它有很多其他优点。条形码、标签序列和阵列都是通用的，在不改变阵列的情况下，条形码可以在不同 SNP 组合的多重反应体系中重复使用。由于阵列上固定的样品点的密度可以非常高，标签阵列结合高多重性的 OLA 反应成为同时扫描大量 SNP 的一种极具吸引力的方法。然而，因为每扫描一个样本都需要一个独立的标签阵列，所以如果很难制作大量的阵列，这种方法将仅限于中等规模样本量的分析。（下面的一段讨论“阵列之阵列”，阐述了这个问题。）

最后，可以用流式微球（或称微珠）分析法（Iannone et al. 2000）。这种固相技术检测 OLA 基因分型产物。可用的微球种类很多，每种微球的红色和黄色荧光比例不同，每种微球可以和一个不同的条形码偶联。OLA 反应体系中，等位基因特异性寡核苷酸均包含一段不同的条形码，位点特异性寡核苷酸均标记荧光素。OLA 反应完成后，连接反应产物与一系列微球杂交。用流式细胞仪检测杂交微球上的荧光信号，根据红色/黄色荧光比值鉴别微球类型（确定 SNP 位点的等位基因），绿色（荧光素）荧光信号提示一个反应中某种 SNP 等位基因的量。这套基因分型系统可以替代基于电泳的检测技术，对于有方便的流式细胞仪使用条件的研究者来说，采用该系统将是一种可行的选择。和基于阵列的技术相比，它适用于各种数量 SNP 和样本的分析，比较灵活。

**引物延伸法**（primer extension assay）

利用引物延伸原理进行等位基因分析的基因分型方法有很多种。单碱基延伸（SBE）或者微型测序法中，位点特异性延伸引物5′端与靶序列退火，3′端结束在SNP位点前一碱基处。在DNA聚合酶作用下延伸，开始延伸的第一个碱基就是多态性碱基（图10-1H）（Syvanen et al. 1990）。反应掺入的核苷酸为双脱氧核苷三磷酸（ddNTP），所以引物延伸一个核苷酸即停止。因为SBE基因分型寡核苷酸序列较短，不需修饰，不需标记，所以价格较低，而且探针的设计也很简单。然而，SBE技术不能用于小的插入或缺失多态性的基因分型。

模板指导的末端碱基掺入终止反应是基于引物延伸原理的最常用的基因分型方法之一，通过荧光偏振技术检测（template-directed primer extension detected by fluorescence polarization，FP-TDI）（Chen et al. 1999），第11章对该技术做了详细介绍。简单地说是指荧光基团受平面偏振光激发后发出偏振荧光。影响荧光偏振度的因素包括反应温度、样本黏度和荧光标记分子的相对分子质量。SBE反应完成后，荧光标记分子明显加重，可以通过荧光偏振的变化检测。FP-TDI在PCR仪上进行基因组DNA扩增，用酶标仪检测。FP-TDI法只能用于单位点检测，但是可以对大样本做筛查。反应体系很容易建立，因为等位基因的识别检测直接在反应板里进行，SBE基因分型探针序列较短，不需修饰，不需标记，价格较低。如果有多重性检测的需要，可用SNaPshot SBE技术（http://www.appliedbiosystems.com），能实现10重反应。多重SNaPshot反应体系包括4种荧光标记的ddNTPs，基因分型探针5′端加一段长度不等的尾序列，引物延伸产物用毛细管测序仪进行鉴定。许多基因分型方法都是建立在SNP是复等位基因的假设基础上，如果某一个体发生第三种等位基因的罕见变异，基因分型结果将是错误的。因为SNaPshot反应体系包括4种ddNTP，所以能准确检测有两种以上等位基因的SNP位点的基因型。

上面提到的这种SBE法的缺点是仅能实现低水平的多重性检测。提高通量的手段之一是在固体基质上进行SBE反应，取代液相反应（如阵列；Pastinen et al. 1997）。简单地讲，即先将位点特异性基因分型寡核苷酸固定在微阵列基质上，然后把多重PCR产物杂交到阵列上。如果探针和PCR产物退火，PCR产物就可以作为SBE反应的模板。鉴于芯片可以容纳数千个不同的寡核苷酸，可能实现复用能力极高的SBE反应。用荧光芯片扫描仪检测每个寡核苷酸处延伸掺入的ddNTPs即可确定每个SNP位点的基因型（Pastinen et al. 1997）。一种类似的固相SBE基因分型检测技术，即流式细胞仪微球分析（Chen et al. 2000），与前面描述的OLA反应的检测方法相似。这些基于阵列和微球的SBE方法的限速步骤可能决定了单个PCR反应的复用水平。

有些课题不需要筛查大量SNPs，高度多重性的SBE基因分型系统就变得不适用了。Pastinen等（2000）采用基于阵列的SBE基因分型策略，并应用了一种“阵列之阵列”的方法。一张微阵列基质被分成多个独立的反应小室，每个小室为一个亚阵列，用于筛查不同个体的同一组SNP，通过减少待检测SNP的数量，可以增大同时筛查的样本量。SNPstream基于阵列的SBE技术（http://www.beckman.com）将阵列之阵

列方法与通用标签阵列（前文）相结合，能够在384孔微量反应板的每个反应孔里实现12重基因分型反应（Bell et al. 2002）。每个反应孔里包括一个小的16点标签阵列（4个对照点，12个标签序列点），印在管底。多重PCR结束后，清除引物及未掺入的核苷酸，在标记ddNTP存在的情况下，在液相中进行SBE反应。12种位点特异性基因分型寡核苷酸5′端标签序列与12种微阵列上固定的标签序列互补。延伸反应产物与阵列上的探针杂交结束后，洗板，除去游离的、未杂交上的物质，扫描反应板，识别每个标签序列的荧光标记。根据荧光信号即可确定12种目标序列的基因型。SNPstream方法的缺点是每个反应里只加了2种ddNTP，即同时检测的SNP位点必须组成相同的等位基因对。将实验设计成检测另一条DNA链上相应位点的基因型可在一定程度上解决这个问题。例如，欲检测A/C SNP，可以设计成检测另一条链上对应的T/G SNP。

与SBE不同，焦磷酸测序引物末端能延伸几个核苷酸（Ronaghi et al. 1996，1998；Alderborn et al. 2000；Langaee and Ronaghi 2005）。原理如下（图10-1I）：测序引物能够与单链DNA模板杂交，在DNA聚合酶的作用下，依次加入4种不同的脱氧核糖核苷酸（dNTP），如果加入的碱基与模板上引物末端对应的下一个碱基互补，则发生延伸反应，释放出焦磷酸（pyrop--hosphatase，PPi），PPi在ATP硫酸化酶的催化作用下转化成ATP。ATP激活荧光素酶发出荧光，产生的荧光强度可以用感光器件定量检测。反复循环加入脱氧核糖核苷酸，延伸测序引物。在进行下一轮延伸之前，必须洗脱样本或者加入核苷酸降解酶，去除前一轮反应中未结合的dNTP和ATP。经循环延伸，即合成一段短DNA序列，可能包含多态SNP或短的插入或缺失多态性片段。尽管焦磷酸测序技术可能做到对多个模板DNA片断同时检测（Pourmand et al. 2002），但通常一次仅分析一个SNP位点。酶催化和底物加入步骤多使这种检测方法更复杂，也增加了成本。而且，必须有专用的焦磷酸测序仪。

## 基因分型平台的应用

不幸的是，没有一种理想的基因分型平台适用于所有课题。研究者应该基于具体项目选择基因分型系统。判定标准见如下几条。

- 需要检测多少个SNPs位点？多少样本？
- 预计项目的总成本是多少？项目计划书里往往只计算检测每个基因型所需耗材的成本，有必要计算耗材、人员和仪器设备的总成本。
- 是否需要特殊仪器？
- 使用商品化基因分型平台还是自己内部的基因分型方法？后者需要做大量的开发和故障排除工作，尤其对于缺乏SNP基因分型经验的实验室。
- 样本筛查模块是固定的还是可变的？对于固定的模块（如作图群体），要对所有个体同时做筛查实验。对于可变DNA筛查模块，连续逐个加样，如法医学实验，每天使用相同的一组标记检测不同个体的基因型。有些基因分型方法很难应用于可变DNA筛查模块。
- 基因组DNA资源是否有限？如果是这样，就有必要选择每个反应只需要很

少量 DNA 的高度多重性检测的方法。

这些问题在下面做具体讨论，经过综合考虑选择的平台可能很适合这 4 种不同规模的基因分析项目。一般情况下，不是每个项目都可以随意选用任何技术，有些技术甚至是不合乎要求的。自己实验室已经建立并且正在使用的基因分型技术可能更可取，即使它并不完全适合预期项目。而且，有无专用仪器也会限制研究者可选择方法的范围。

## 低通量：10 个 SNP 位点-100 个样本

小项目可能作为一个大的研究项目的前期工作，或者研究者只对几个有潜在功能的 SNP 及它们在一两个群体的小样本中的频率感兴趣（如 Wilson et al. 2001）。对于一个小项目，没有必要使用能够高多重性的基因分型方法，除非基因组 DNA 资源非常有限。然而，购买寡核苷酸或探针的成本是需要考虑的主要问题，即检测方法的固定成本。这个成本将分配到每个待分型样本中去，固定检测成本高的方法（如 TaqMan）对于小样本量的研究不够经济有效。对于 10 个 SNP 位点 100 个样本这种通量的实验，RFLP 分析、等位基因特异性 PCR，或者基于 SBE 的 FP-TDI 等单重基因分型方法可能是最有效的。如果具备必要的仪器设备，也可以用焦磷酸测序技术。例如，De Luca 等（2003）对 173 个果蝇自交系 12 个 SNP 位点和小插入/缺失事件基因分型，以识别衰老相关变异的多态性位点。假如选用每个检测反应固定成本低的基因分型系统，低通量项目的成本可能主要体现在劳动力上，尽管该成本可能也相对较低。

## 中等通量：10 个 SNP 位点-1000 个样本

该通量的项目以 follow-up 的实验策略，达到大范围关联定位扫描的目的，即在第二个大样本中复制一小组潜在的可能关联的 SNP 位点（Genissel et al. 2004；Smyth et al. 2004；Dworkin et al. 2005）。样本量大，检测寡核苷酸/探针的成本就成为次要问题了，因为这笔成本在大量样本间按比例分配。因为待分析的 SNP 数量少，使用非多重的检测方法是可行的，尽管这样要做 10 000 个独立的基因分型反应，消耗品的成本也很高。如果有一台通量足够高的基因扩增仪和一台定量 PCR 仪或者荧光酶标仪，TaqMan 法、Invasive 分析法，或者 FP-TDI SBE 法均可选用。这些反应体系很容易建立，即便没有先进的液体处理系统，在基因分型反应板里直接进行等位基因检测也可以。根据片段长度或荧光标记的不同，用毛细管测序的方法（如 SNaPshot）处理等位基因特异性产物是可能的，但是即使用 10 重的反应，也要跑 1000 条电泳，也许对于许多实验室来讲，都不是高效率的、经济有效的方法。可替代的方法是尼龙膜上基于 ASO 的检测方法，将 PCR 扩增产物印在膜上，与同位素标记的探针杂交。例如，Zimmerman 等（2000）用 ASO 法对 800 个果蝇 16 个 SNP 作基因分型，对控制翅膀形状的数量性状位点进行定位。用这种方法，可以在膜上转印许多样本，能提高大样本项目的通量，尽管需要优化每个探针的杂交条件，只能检测相对较少的 SNPs。要得到 10 个 SNP 1000 个样本的基因分型数据，基于阵列的 ASO 法是一种非常廉价的方法，但是，与完全在液相中进行的方法（如 FP-TDI）相比，需要的劳动力强度可能更高。另外需要注意的是，基于阵列的 ASO 法适用于同时筛查的样本数固定的筛查模块，如果以后

用同样的方法做一些额外样本的基因分型，效率将非常低。

## 中等通量：100 个 SNP 位点-100 个样本

这种项目通量向研究者提出了挑战，是从事全基因组 QTL 定位或 SNP 连锁图构建的研究者经常遇到的问题。每个检测反应的固有成本必须低（因为样本数量少），但是需要复用能力（因为待检测 SNP 数量多）。有些复用方法，如基于 OLA 的 SNPWave 技术，是一种适合 100 个 SNP 位点-100 个样本这种通量项目的比较突出的方法。例如，利用 SNPWave 技术 100 重反应对 92 个品种拟南芥的 100 个 SNP 位点基因分型，定位开花时间性状的 QTL（El-Lithy et al. 2006）。高度复用确保了基因分型工作简单化，只需要在毛细管测序仪上跑少量电泳即可。唯一的顾虑是基因分型探针长，则每个基因型需要的费用就高。基于 SBE 的 SNPstream 系统是一种值得选择的方法。该系统中，反应过程和基因型检测直接在微量反应板里进行，SNPstream 能达到 12 重反应能力。尽管该技术适合这种通量的基因分型项目，但是要求具备适合的仪器。如果研究者有专业的单点（或低多重性）基因分型方法的经验，又具备合适的仪器设备，不用多重基因分型反应，也可以完成几百个人 100 个 SNP 的基因分型工作。例如，Smith 等（2005）采用 SBE 法、RFLP 和等位基因特异性 PCR 等技术对大约 100 个样本 500 个 SNP 作基因分型，构建了蝾螈类美西螈属连锁图。

## 高通量：100 个 SNP 位点-1000 个样本

这是在一般水平的实验室能进行的最大通量的项目，唯一需要保障的是新仪器设备的支出。如果要进行更大通量的实验，许多研究者情愿将基因分型工作外包给基因组中心或公司去做，和（或）采用第 13 章介绍的超高通量的技术。使用介绍过的任意一种方法都能得到 100 000 个基因型，但一般说来，那些多重性高的技术可能效率最高。在毛细管测序仪上，通过等位基因特异性基因分型产物片段大小确定基因型是完全可能的。然而，即使用 100 重的反应（如 SNPWave），还需要跑 1000 条电泳，并且要计数每个条带产物峰的有无、片段大小、峰的高度和荧光基团——100 重检测实验仍是劳动密集型的。此外，在实际中，还可能达不到这么高的多重水平，就需要跑更多的电泳，消耗品的成本也会随之按比例增加。因此，能同时筛查大量样本、基因型检测更加自动化的方法更可取，那就是基于阵列或基于微量反应板的系统。例如，Macdonald 等（2005a）采用 16 重基于 OLA 的基因分型技术，通过同位素标记探针和基于阵列的等位基因检测技术，得到 2000 个果蝇 200 个 SNP 和插入/缺失多态性位点的基因型数据。这个方法消耗品成本低廉，也不需要非常专业的仪器设备，但是有些实验步骤是劳动力密集型的。另外，这种基于阵列的方法最适合样本数固定的筛查模块；如果有额外的样本不断加进来，基于阵列的方法就显得不实用了。

## 多态性的问题

SNP 是个体间序列水平变异最丰富的形式。在靶 SNP 周围或基因分型探针的结合区域内，还会有另外的、未知的、分离的多态性位点。如果是这样的话，基因分型寡核

苷酸会与这个次要 SNP 的一种特异性等位基因互补。因此，如果某样本该位点处是另外一种（破坏性）等位基因，杂交探针与靶序列形成的复合体不够稳定（Wallace et al. 1979），对靶 SNP 等位基因的检测就失败了（图 10-3）。如果某样本的破坏性次要 SNP 位点等位基因是一对纯合子，检测实验将完全失败，得不到靶 SNP 位点的基因型。更严重的是，如果次要 SNP 位点等位基因是一对杂合子，就会只漏掉靶 SNP 位点的一个等位基因，杂合子将被误判为纯合子（图 10-3）。

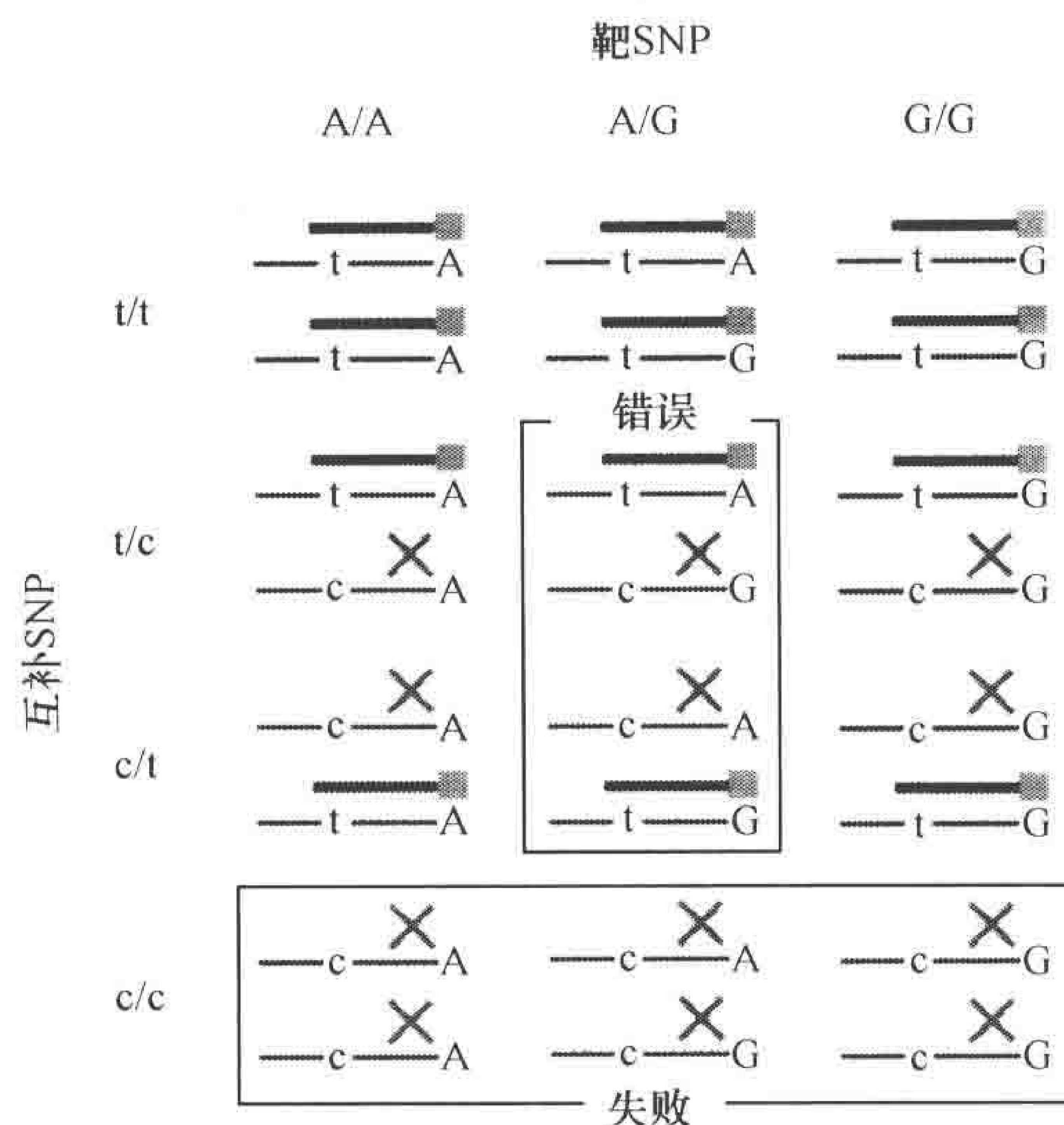



图 10-3 基因分型过程中出现其他多态性位点的问题。本例中，用 SBE 法检测等位基因为 A/G 的靶 SNP，基因分型寡核苷酸结合区内存在另外一个等位基因为 t/c 的 SNP。基因分型寡核苷酸被设计成与次要 SNP“t”等位基因结合，而“c”等位基因完全破坏寡核苷酸结合。如果一个样本靶 SNP 和次要 SNP 都是杂合子，靶 SNP 的基因型将被错误地报告为纯合子。此外如果次要 SNP 是等位基因“c”的纯合子，基因分型检测将失败

出现次要 SNPs，即使不是所有方法，也是许多基因分型方法都会遇到的问题（Matsuzaki et al. 2004；Macdonald et al. 2005b；Koboldt et al. 2006），尽管问题的程度有平台特异性（Koboldt et al. 2006）。用于基因分型的生物种类也是一个重要的影响因素。核苷酸多样性低的生物（如人类）与核苷酸多样性更高的生物（如果蝇）相比，受到次要等位基因的影响要小很多，因为探针结合区存在另一 SNP 的可能性相对要小一些。对多个样本的靶序列测序能确定大多数常见多态性，附近没有其他 SNP 的位点即可进行基因分型。通过设计简并基因分型寡核苷酸可能消除已知的次要 SNP 的干扰（Macdonald et al. 2005b），尽管不知道这种处理是否对其他基因分型平台有效。或者，选用一种对等位基因特异性探针的定位比较灵活的基因分型平台。例如，按另外一条 DNA 链设计探针（如 SBE 法），或者稍微移动一下等位基因特异性探针相对于靶 SNP 的位置（如 ASO，TaqMan）。然而，经过测序适当数量样本，如果没有发现其他 SNP，几乎就不会有破坏性 SNP 存在的可能（Macdonald et al. 2005b）。

# 参考文献


Alderborn A., Kristofferson A., and Hammerling U. 2000. Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. *Genome Res.* **10:** 1249–1258.

Banér J., Isaksson A., Waldenstrom E., Jarvius J., Landegren U., and Nilsson M. 2003. Parallel gene analysis with allele-specific padlock probes and tag microarrays. *Nucleic Acids Res.* **31:** e103.

Barany F. 1991. Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proc. Natl. Acad. Sci.* **88:** 189–193.

Bell P.A., Chaturvedi S., Gelfand C.A., Huang C.Y., Kochersperger M., Kopla R., Modica F., Pohl M., Varde S., Zhao R., et al. 2002. SNPstream UHT: Ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. *Biotechniques* **Suppl:** 70–77.

Bottema C.D. and Sommer S.S. 1993. PCR amplification of specific alleles: Rapid detection of known mutations and polymorphisms. *Mutat. Res.* **288:** 93–102.

Chen J., Iannone M.A., Li M.S., Taylor J.D., Rivers P., Nelsen A.J., Slentz-Kesler K.A., Roses A., and Weiner M.P. 2000. A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Res.* **10:** 549–557.

Chen X., Levine L., and Kwok P.Y. 1999. Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res.* **9:** 492–498.

Conner B.J., Reyes A.A., Morin C., Itakura K., Teplitz R.L., and Wallace R.B. 1983. Detection of sickle cell $\beta^S$-globin allele by hybridization with synthetic oligonucleotides. *Proc. Natl. Acad. Sci.* **80:** 278–282.

Day D.J., Speiser P.W., White P.C., and Barany F. 1995. Detection of steroid 21-hydroxylase alleles using gene-specific PCR and a multiplexed ligation detection reaction. *Genomics* **29:** 152–162.

De La Vega F.M., Lazaruk K.D., Rhodes M.D., and Wenz M.H. 2005. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan® SNP genotyping assays and the SNPlex™ genotyping system. *Mutat. Res.* **573:** 111–135.

De Luca M., Roshina N.V., Geiger-Thornsberry G.L., Lyman R.F., Pasyukova E.G., and Mackay T.F. 2003. Dopa decarboxylase (*Ddc*) affects variation in *Drosophila* longevity. *Nat. Genet.* **34:** 429–433.

Dworkin I., Palsson A., and Gibson G. 2005. Replication of an *Egfr*-wing shape association in a wild-caught cohort of *Drosophila melanogaster*. *Genetics* **169:** 2115–2125.

Eggerding F.A. 1995. A one-step coupled amplification and oligonucleotide ligation procedure for multiplex genetic typing. *PCR Methods Appl.* **4:** 337–345.

El-Lithy M.E., Bentsink L., Hanhart C.J., Ruys G.J., Rovito D., Broekhof J.L., van der Poel H.J., van Eijk M.J., Vreugdenhil D., and Koornneef M. 2006. New *Arabidopsis* recombinant inbred line populations genotyped using SNPWave and their use for mapping flowering-time quantitative trait loci. *Genetics* **172:** 1867–1876.

Genissel A., Pastinen T., Dowell A., Mackay T.F.C., and Long A.D. 2004. No evidence for an association between common nonsynonymous polymorphisms in *Delta* and bristle number variation in natural and laboratory populations of *Drosophila melanogaster*. *Genetics* **166:** 291–306.

Germer S. and Higuchi R. 1999. Single-tube genotyping without oligonucleotide probes. *Genome Res.* **9:** 72–78.

Gerry N.P., Witowski N.E., Day J., Hammer R.P., Barany G., and Barany F. 1999. Universal DNA microarray method for multiplex detection of low abundance point mutations. *J. Mol. Biol.* **292:** 251–262.

Grossman P.D., Bloch W., Brinson E., Chang C.C., Eggerding F.A., Fung S., Iovannisci D.M., Woo S., and Winn-Deen E.S. 1994. High-density multiplex detection of nucleic acid sequences: Oligonucleotide ligation assay and sequence-coded separation. *Nucleic Acids Res.* **22:** 4527–4534.

Hall J.G., Eis P.S., Law S.M., Reynaldo L.P., Prudent J.R., Marshall D.J., Allawi H.T., Mast A.L., Dahlberg J.E., Kwiatkowski R.W., et al. 2000. Sensitive detection of DNA polymorphisms by the serial invasive signal amplification reaction. *Proc. Natl. Acad. Sci.* **97:** 8272–8277.

Howell W.M., Jobs M., Gyllensten U., and Brookes A.J. 1999. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. *Nat. Biotechnol.* **17:** 87–88.

Iannone M.A., Taylor J.D., Chen J., Li M.S., Rivers P., Slentz-Kesler K.A., and Weiner M.P. 2000. Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry. *Cytometry* **39:** 131–140.

Jobs M., Howell W.M., Stromqvist L., Mayr T., and Brookes A.J. 2003. DASH-2: flexible, low-cost, and high-throughput SNP genotyping by dynamic allele-specific hybridization on membrane arrays. *Genome Res.* **13:** 916–924.

Koboldt D.C., Miller R.D., and Kwok P.Y. 2006. Distribution of human SNPs and its effect on high-throughput genotyping. *Hum. Mutat.* **27:** 249–254.

Landegren U., Kaiser R., Sanders J., and Hood L. 1988. A ligase-mediated gene detection technique. *Science* **241:** 1077–1080.

Langaee T. and Ronaghi M. 2005. Genetic variation analyses by Pyrosequencing. *Mutat. Res.* **573:** 96–102.

Lazzaro B.P., Sceurman B.K., Carney S.L., and Clark A.G. 2002. fRFLP and fAFLP: Medium-throughput genotyping by fluorescently post-labeling restriction digestion. *Biotechniques* **33:** 539–546.

Li H., Cui X., and Arnheim N. 1990. Direct electrophoretic detection of the allelic state of single DNA molecules in human sperm by using the polymerase chain reaction. *Proc. Natl. Acad. Sci.* **87:** 4580–4584.

Livak K.J. 1999. Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. *Genet. Anal. Biomol. Eng.* **14:** 143–149.

Livak K.J., Flood S.J., Marmaro J., Giusti W., and Deetz K. 1995. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl.* **4:** 357–362.

Lu M., Shortreed M.R., Hall J.G., Wang L., Berggren T., Stevens P.W., Kelso D.M., Lyamichev V., Neri B., and Smith L.M. 2002. A surface invasive cleavage assay for highly parallel SNP analysis. *Hum. Mutat.* **19:** 416–422.

Luo J., Bergstrom D.E., and Barany F. 1996. Improving the fidelity of *Thermus thermophilus* DNA ligase. *Nucleic Acids Res.* **24:** 3071–3078.

Lyamichev V., Mast A.L., Hall J.G., Prudent J.R., Kaiser M.W., Takova T., Kwiatkowski R.W., Sander T.J., de Arruda M., Arco D.A., et al. 1999. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat. Biotechnol.* **17:** 292–296.

Macdonald S.J., Pastinen T., and Long A.D. 2005a. The effect of polymorphisms in the enhancer of split gene complex on bristle number variation in a large wild-caught cohort of *Drosophila melanogaster*. *Genetics* **171:** 1741–1756.

Macdonald S.J., Pastinen T., Genissel A., Cornforth T.W., and Long A.D. 2005b. A low-cost open-source SNP genotyping platform for association mapping applications. *Genome Biol.* **6:** R105.

Matsuzaki H., Loi H., Dong S., Tsai Y.Y., Fang J., Law J., Di X., Liu W.M., Yang G., Liu G., et al. 2004. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14:** 414–425.

Newton C.R., Graham A., Heptinstall L.E., Powell S.J., Summers C., Kalshekerl N., Smith J.C., and Markham A.F. 1989. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Res.* **17:** 2503–2516.

Nickerson D.A., Kaiser R., Lappin S., Stewart J., Hood L., and Landegren U. 1990. Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. *Proc. Natl. Acad. Sci.* **87:** 8923–8927.

Nilsson M., Malmgren H., Samiotaki M., Kwiatkowski M., Chowdhary B.P., and Landegren U. 1994. Padlock probes: Circularizing oligonucleotides for localized DNA detection. *Science* **265:** 2085–2088.

Olivier M. 2005. The Invader assay for SNP genotyping. *Mutat. Res.* **573:** 103–110.

Pastinen T., Kurg A., Metspalu A., Peltonen L., and Syvänen A.C. 1997. Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.* **7:** 606–614.

Pastinen T., Raitio M., Lindroos K., Tainola P., Peltonen L., and Syvänen A.C. 2000. A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res.* **10:** 1031–1042.

Pourmand N., Elahi E., Davis R.W., and Ronaghi M. 2002. Multiplex Pyrosequencing. *Nucleic Acids Res.* **30:** e31.

Prince J.A., Feuk L., Howell W.M., Jobs M., Emahazion T., Blennow K., and Brookes A.J. 2001. Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): Design criteria and assay validation. *Genome Res.* **11:** 152–162.

Rao K.V., Stevens P.W., Hall J.G., Lyamichev V., Neri B.P., and Kelso D.M. 2003. Genotyping single nucleotide polymorphisms directly from genomic DNA by invasive cleavage reaction on microspheres. *Nucleic Acids Res.* **31:** e66.

Ririe K.M., Rasmussen R.P., and Wittwer C.T. 1997. Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Anal. Biochem.* **15:** 154–160.

Ronaghi M., Uhlen M., and Nyren P. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281:** 363–365.

Ronaghi M., Karamohamed S., Pettersson B., Uhlen M., and Nyren P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242:** 84–89.

Saiki R.K., Bugawan T.L., Horn G.T., Mullis K.B., and Erlich H.A. 1986. Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. *Nature* **324:** 163–166.

Saiki R.K., Scharf S., Faloona F., Mullis K.B., Horn G.T., Erlich H.A., and Arnheim N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230:** 1350–1354.

Schouten J.P., McElgunn C.J., Waaijer R., Zwijnenburg D., Diepvens F., and Pals G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30:** e57.

Smith J.J., Kump D.K., Walker J.A., Parichy D.M., and Voss S.R. 2005. A comprehensive expressed sequence tag linkage map for tiger salamander and Mexican axolotl: Enabling gene mapping and comparative genomics in *Ambystoma*. *Genetics* **171:** 1161–1171.

Smyth D., Cooper J.D., Collins J.E., Heward J.M., Frankly J.A., Howson J.M.M., Vella A., Nutland S., Rance H.E., Maier L., et al. 2004. Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes* **53:** 3020–3023.

Syvänen A.C., Aalto-Setala K., Harju L., Kontula K., and Söderlund H. 1990. A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8:** 684–692.

Thomas M.G., Bradman N., and Flinn H.M. 1999. High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum. Genet.* **105:** 577–581.

Tobe V.O., Taylor S.L., and Nickerson D.A. 1996. Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. *Nucleic Acids Res.* **24:** 3728–3732.

van Eijk M.J.T., Broekhof J.L.N., van der Poel H.J.A., Hogers R.C.J., Schneiders H., Kamerbeek J., Verstege E., van Aart J.W., Geerlings H., Buntjer J.B., et al. 2004. SNPWave™: A flexible multiplexed SNP genotyping technology. *Nucleic Acids Res.* **32:** e47.

Vos P., Hogers R., Bleeker M., Reijans M., van de Lee T., Hornes M., Frijters A., Pot J., Peleman J., Kuiper M., and Zabeau M. 1995. AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Res.* **23:** 4407–4414.

Wallace R.B., Shaffer J., Murphy R.F., Bonner J., Hirose T., and Itakura K. 1979. Hybridization of synthetic oligodeoxyribonucleotides to φχ174 DNA: The effect of single base pair mismatch. *Nucleic Acids Res.* **6:** 3543–3557.

Wang J., Chuang K., Ahluwalia M., Patel S., Umblas N., Mirel D., Higuchi R., and Germer S. 2005. High-throughput SNP genotyping by single-tube PCR with $T_m$-shift primers. *Biotechniques* **39:** 885–893.

Wilson J.F., Weale M.E., Smith A.C., Gratrix F., Fletcher B., Thomas M.G., Bradman N., and Goldstein D.B. 2001. Population genetic structure of variable drug response. *Nat. Genet.* **29:** 265–269.

Ye S., Dhillon S., Ke X., Collins A.R., and Day I.N. 2001. An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res.* **29:** E88.

Zhang R., Zhu Z., Zhu H., Nguyen T., Yao F., Xia K., Liang D., and Liu C. 2005. SNP Cutter: A comprehensive tool for SNP PCR-RFLP assay design. *Nucleic Acids Res.* **33:** W489–492.

Zimmerman E., Palsson A., and Gibson G. 2000. Quantitative trait loci affecting components of wing shape in *Drosophila melanogaster*. *Genetics* **155:** 671–683.

# 互联网资源

**http://bioinfo.bsd.uchicago.edu/SNP_cutter.htm** SNP Cutter: SNP PCR-RFLP Assay Design, developed by Zhang et al., University of Chicago, Illinois.

# 11 实验室用中通量基因分型的实验策略

Edwin Cuppen,[1] Stuart J. Macdonald,[2] Connie Ha,[3] Pui-Yan Kwok,[4] W. Brad Barbazuk,[5] An-Ping Hsia,[6] Hsin D. Chen,[6] Yan Fu,[5] Kazuhiro Ohtsu,[6] and Patrick S. Schnable[6,7,8]

*[1]Hubrecht Laboratory, Utrecht, The Netherlands; [2]Department of Ecology and Evolutionary Biology and Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66045; [3]Cardiovascular Research Institute and Center for Human Genetics and [4]Department of Dermatology, University of California, San Francisco, California 94143-0793; [5]Donald Danforth Plant Science Center, St. Louis, Missouri 63132; [6]Department of Agronomy, [7]Department of Genetics, Development, and Cell Biology, and [8]Center for Plant Genomics, Iowa State University, Ames, Iowa 50011*



## 简介

在本书的第 9 章，我们讨论了各种实验室通量的基因分型平台和限于中等通量的基因分型的分析策略。这一章我们将提供一系列具体的基因分型实验设计，这些设计描述了适合处理几千到几十万基因分型的分析途径。

## 方案一 等位基因扩增法进行基因分型

Edwin Cuppen

*Hubrecht Laboratory, Utrecht, The Netherlands*

本方案描述了用荧光检测（KASPar）扩增的特殊等位基因的基因分型方法。这种操作步骤得到的结果与任何一种实时 PCR 的方案结果是一致的，并且具有无需对寡核苷酸特殊标记的优点。更多信息，登录 http://www.kbioscience.co.uk。

# 材料

## 试剂

基因组模板 DNA
$MgCl_2$（50mmol/L）
反应混合物（KASPar，Kbioscience）
Taq 聚合酶（Kbioscience）

## 仪器

荧光读板或者实时 PCR 仪（与 FAM、VIC 和 ROX 相匹配）
Klustercaller 软件
PCR 仪
反应板

# 方法

（1）以 Web 为基础设计寡核苷酸链（2 个 40mer 等位基因特异和一个 20mer 的共同等位基因）：http://www.kbioscience.co.uk/primer-picker/。

（2）制备反应混合物（总体积 4μL）。

基因组模板 DNA（10～100ng）　2μL
反应混合物（KASPar）　1μL
寡核苷酸混合物　0.055μL（每个等位基因特异寡核苷酸是 12μmol/L，共同等位基因的寡核苷酸是 30μmol/L）
Taq 聚合酶　0.013μL
$MgCl_2$（50mmol/L）　0.032μL
MilliQ $H_2O$　0.9μL

（3）（按下列方案进行）PCR。

94℃　15min；
94℃　10s，57℃　5s，72℃10s，20 个循环；
94℃　10s，57℃　20s，72℃　40s，18 个循环。

（4）荧光读板或者 实时 PCR 仪（与 FAM、VIC 和 ROX 相匹配）扫描板。

（5）Klustercaller 软件分析数据。

## 方案二　双脱氧重测序法进行基因分型

Edwin Cuppen
*Hubrecht Laboratory, Utrecht, The Netherlands*

本实验描述了 PCR 片段的双脱氧重测序法，该操作方法是一种快速灵活的基因分

型技术。这种策略能辨别并确定基因组内感兴趣区域杂合子 SNP 的存在。序列数据可用各种软件包进行分析，包括 Staden（Bonfield et al. 1998）或者 Polyphred（Stephenset al. 2006）。

# 材料

注意：请参照附件正确处理标记〈!〉的材料。

## 试剂

琼脂糖凝胶（1%）
BigDye 稀释缓冲液（2.5×，Applied Biosystems）
BigDye 终止试剂 v3.0（Applied Biosystems）
dNTP（每个 10 mmol/L）
乙醇（80%）
甲酰胺〈!〉(Optional，see Step 16)
Forward 引物（2 mmol/L）
基因组模板 DNA
PCR 缓冲液（10×，来自 Taq 酶的供应商）
沉淀混合物 1L 中混合 800 mL 96%乙醇，16 mL 的乙酸钠〈!〉（pH 5.5）和 158 mL MilliQ $H_2O$
反向引物（2 mmol/L）
序列引物（2 mmol/L；用作 PCR 扩增的寡核苷酸之一）
Taq 聚合酶（5U/μL）

## 仪器

琼脂糖电泳仪器
毛细管测序仪（如 Applied Biosystems3730 标准的 RapidSeq 实验设计）
电热板预设至 80℃［可选做，步骤（15)］
PCR 仪
反应板或者反应管
序列分析软件（如 Staden 或者 Polyphred）
涡旋搅拌器

# 方法

(1) 设计一个包含多态位点的基因组片段的扩增子进行 PCR 扩增。PCR 和测序的最佳大小约为 300bp。

(2) 制备 PCR 反应体系（总体积 10μL)。

| | |
|---|---|
| 基因组模板 DNA（10～100ng） | 5μL |
| 10×PCR 缓冲液 | 1μL |
| 正向引物（2 mmol/L） | 1μL |
| 反向引物（2 mmol/L） | 1μL |
| dNTP（每个 10 mmol/L） | 0.4μL |
| Taq 聚合酶（5U/μL） | 0.4μL |
| MilliQ $H_2O$ | 1.2μL |

（3）用下面的循环退火温度降落实验设计进行 PCR：

94℃ 60s；

94℃ 20s，65℃ 20s 并且每一次循环递减 0.6℃，72℃ 30s，12 个循环；

接下来的 20 个循环 92℃ 20s，58℃ 20s，72℃ 30s；

72℃ 180s。

（4）取 PCR 产物 1～2μL 在 1%琼脂糖上进行检验。如果见一清晰的条带，则 PCR 产物无需纯化即可进行测序。否则，优化 PCR 条件，如进行梯度 PCR。利用琼脂糖凝胶切除片段或者经纯化的成效不高 PCR 弱产物进行序列分析是可能的，但建议不要这样做。

（5）用 25μL MilliQ $H_2O$ 稀释剩下的 PCR 产物。

（6）制备双脱氧重测序反应体系（总体积 5μL）。

| | |
|---|---|
| 稀释后的 PCR 产物 | 1μL |
| 测序引物（2 mmol/L） | 1μL |
| BigDye 终止试剂 v3.0 | 0.2μL |
| 2.5×BigDye 稀释缓冲液 | 1.8μL |
| MilliQ $H_2O$ | 1μL |

（7）循环测序实验。

92℃ 10s，50℃ 5s，60℃ 120s，40 个循环

（8）加入 30μL 沉淀混合物来纯化产物。

（9）将上述混合物振荡 15s。

（10）3200$g$ 离心 40min（板）或者 10 000$g$ 离心 10min（管）。

（11）弃上清（反应板以 32$g$ 倒置离心 1min）。

（12）加 25μL 80%的乙醇洗涤沉淀。

（13）3200$g$ 离心 5min（板）或者 10 000$g$ 离心 2min（管）。

（14）弃上清（反应板以 32$g$ 颠倒离心 1min）。

（15）空气中晾干沉淀或者 80℃加热 10～15min（直至无乙醇味，不可过干，避光）。

（16）用 10μL 水或者甲酰胺溶解沉淀。

（17）毛细管测序仪分析（如 Applied Biosystems 3730）。

（18）分析测序数据，可用多种软件包（如 Staden 或者 Polyphred）。

## 方案三　寡核苷酸连接测定法

Stuart J. Macdonald

*Department of Ecology and Evolutionary Biology and Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66045*

该方案描述了寡核苷酸连接测定法（OLA），该法用含 3 个核苷酸的寡核苷酸链接和耐热 Taq DNA 连接酶来区分单核苷酸多态性（SNP）等位基因。进行 16-plex 的 OLA 基因分型反应，并在膜阵列上用标记的探针检测等位基因特异性的 OLA 产物。图 11-1 给出了基因分型流水线作业的概况。

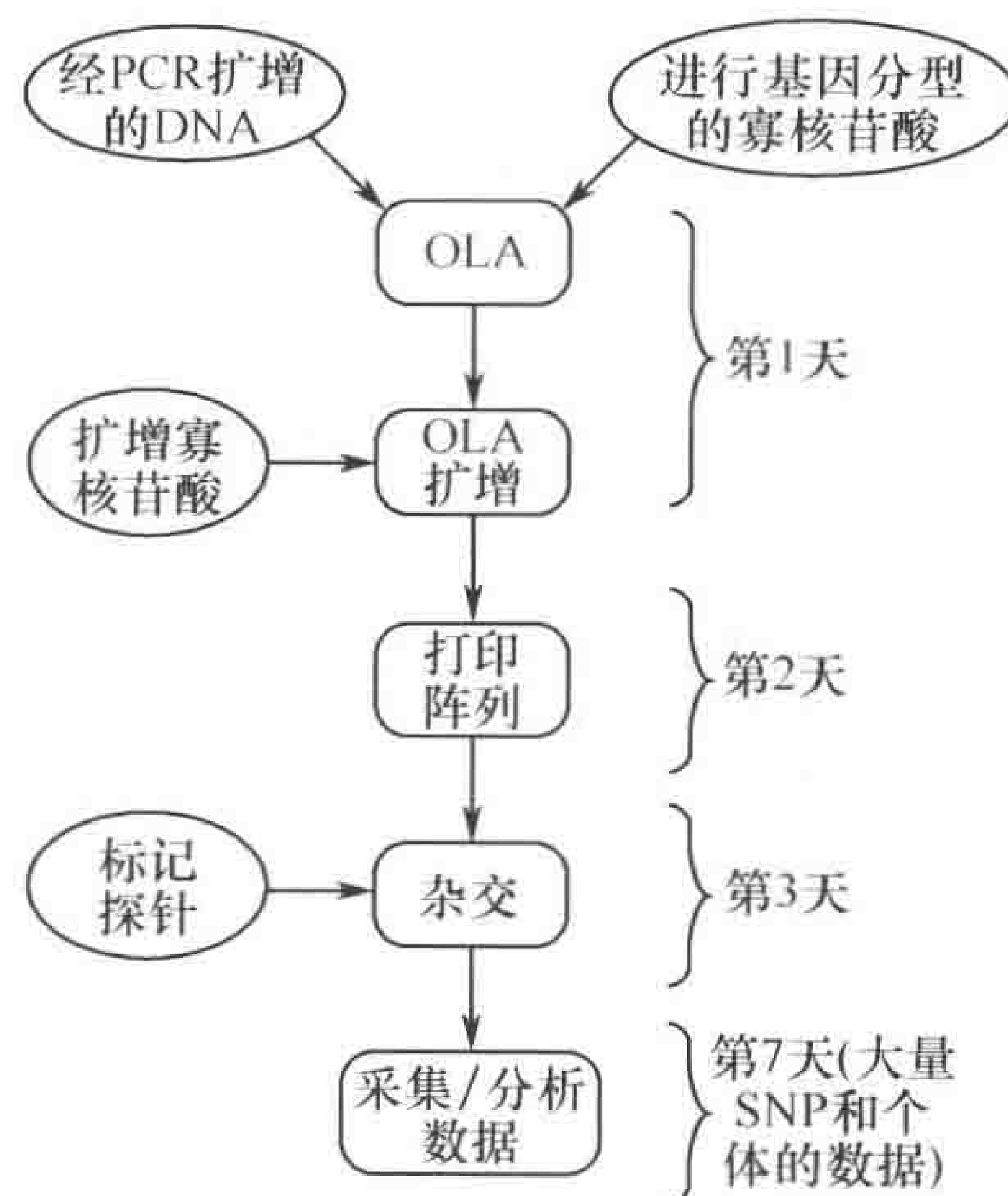



图 11-1　基因分型流水线

## 材料

注意：请参照附件正确处理标记〈!〉的材料。

### 试剂

5′-腺苷三磷酸（ATP）溶液（100 mmol/L；GE Healthcare）

[$\gamma$-$^{33}$P] 5′-ATP（250μCi；PerkinElmer）

10×扩增缓冲液（500 mmol/L KCl，1% Triton X-100），用最小量的扩增缓冲液，因为扩增试剂是直接加在 OLA 反应体系中的，而 OLA 反应体系在 PCR 时已经含有适量浓度的 Tris-HCl 和 $MgCl_2$

变性缓冲液（0.5mol/L NaOH〈!〉，1.5mol/L NaCl）

包含目标 SNP 的 DNA（PCR 已经扩增好）

dNTP（25mmol/L）

DTT（25mmol/L）〈!〉

鲱鱼精子 DNA，超声处理（10μg/μL；Promega）

杂交缓冲液（预温至 42℃）

0.525 mol/L 磷酸钠（pH 7.2）〈!〉

7% SDS〈!〉

1 mmol/L EDTA（pH 8.0）

10 mg/mL 牛血清白蛋白（BSA），每次杂交前配制新鲜的液体并经无菌过滤以除去可能结合在尼龙膜上的杂质。

中性缓冲液（0.4 mol/L Tris-HCl [pH 7.4]，2×SSC）

10 倍 OLA 缓冲液（500 mmol/L Tris-HCl [pH 8.5]，500 mmol/L KCl，75 mmol/L $MgCl_2$，10 mmol/L β-NAD）

寡核苷酸链：以最低体系合成用于基因分型的寡核苷酸链（未经修饰的），表 11-1 为探针和条形码。

表 11-1 为探针和条形码。

**表 11-1　探针和条形码**

| 探针号 | 探针/条形码 a | 探针/条形码 b |
|---|---|---|
| 01 | AT ATTCTGAGACACGCCGCG | AT ACGCGATGGGATCAGACT |
| 02 | AT GCGACTCTTGACGAACGT | TT CGAGCGTCTGGCACACTT |
| 03 | GT CACTCGTGTCCAGGATGT | TA TCGCGTGTCAGTGCTTGT |
| 04 | GA TACCGGACCATGTTTCGC | GA TGTTCGTCCATGCGACCT |
| 05 | TG ATCCGCGTCGATGCTCTT | GC AGTCACGTTCTCGAATCG |
| 06 | TT TAGCCGGATCACCGTGTG | AT ATGTGCAGAACCCGCGAC |
| 07 | AG AGAGACGTTGCCCAAGTC | GA TGCGATACCCTGCGATCT |
| 08 | AT TTAGCGTGCAGCCGACCT | AT GCGTGGTGTCCGATCATA |
| 09 | TA AGGGTTACGAACATCGCC | TG GACTCTCATAACGGCGTC |
| 10 | GC AGCTCGTCACAGGTATTG | TA CCGGATTACAGCTCGTGG |
| 11 | AG CTAATGTCGAGTCACGCT | TC TACACGAGAACGAGGCAC |
| 12 | AG CGCGACGTTGATCCAGAT | AA TGAACGAGACCGCGTGAC |
| 13 | TC GGACTCGTGACGCTATTT | AT GAGAGTTCGATGACCTGT |
| 14 | AC GCACTGACGATCATTCGG | TT CGACCCGGACGACTGAAT |
| 15 | TA TAGCCGTGAACCCGATGC | TA AAGCACAGTCCGTAATCT |
| 16 | AT CATGTCCCAAGCGCGGTA | AA GCCGATGTCGATCTACCT |

注：所有 20-nt 的探针序列方向均是 5′→3′，掺入 OLA 寡核苷酸上游的 16-nt 条形码序列是画线部分的互补序列。

M13 扩增寡核苷酸

M13F. BRL，CCCAGTCACGACGTTGTAAAACG

M13R. BRL，AGCGGATAACAATTTCACACAGG

洗脱缓冲液（0.1%SDS〈!〉），预热至 80℃

T4 多核苷酸激酶（10 units/μL；New England Biolabs），缓冲液由厂家提供

Taq DNA 连接酶（40 units/μL；New England Biolabs），用 10×OLA 缓冲液

Taq DNA 聚合酶（5 units/μL；New England Biolabs），用 10×扩增缓冲液，因为扩增产物非常小，几乎任何一种耐热 DNA 酶都能用

漂洗缓冲液（5×SSPE，0.1%SDS〈!〉），预温至 40℃

### 器材

尼龙膜芯片（定制登录 http://cstern.bio.uci.edu/tools/genotyping.htm）

从免费统计程序包 R（http://www.r-project.org/）获得 Custom 软件（http://cstern.bio.uci.edu/tools/genotyping.htm）

电热板预设至 37℃、65℃、80℃、96℃

杂交炉预设至 42℃

杂交管

图像采集软件包（如 GE Healthcare Array Vision）

实验手套（无粉）

液体处理系统（如 Art Robbins Instruments Hydra）[可选做，见步骤（8）]

手工并行输入工具 [可选做，见步骤（8），（14）]

PCR 板（96 孔或者 384 孔）

PCR 热循环仪（如 Applied Biosystems 双模板 384 孔 9700）

荧光图像仪（如 GE Healthcare Typhoon）

放射性废料收集器

重复移液器 [可选做，见步骤（8）]

摇摆台

存储荧光屏（如 GE Healthcare）

紫外线（UV）光源（50mJ）〈!〉

水浴预设至 40℃、80℃

## 方法

### 制备 OLA 基因分型寡核苷酸链

（1）用序列对比法或者摘录网上数据确定 SNPs。

（2）按下列步骤为每个 SNP 设计两个上游等位基因特异寡核苷酸 [47 个核苷酸（nt）] 和一个公共下游寡核苷酸（31nt）。

上游 _ a，M13F+C+Barcode _ a+Up _ flank+Allele _ a

上游 _ b，M13F+C+Barcode _ b+Up _ flank+Allele _ b

下游，Down _ flank+G+ M13R. RC

OLA 基因分型寡核苷酸的部分是

M13F，GACGTTGTAAAACG

M13R. RC，CCTGTGTGAAATTG

Barcode _ a 和 Barcode _ b 是 16 nt 的条形码序列，该序列允许在杂交过程中区分等位基因（表 11-1）。16 对，允许在 16-plex OLA 的反应中每个 SNP 有一对条形码或探针。

Up _ flank 是靶 SNP 的上游区域特有的 15 个 nt，Down _ flank 是靶 SNP 的下游区域特有的 16 个 nt。如果这两个侧翼区中的任一个能分离一个附加的 SNP，那么这些位点的寡核苷酸序列能够掺入一个简并碱基。

Allele _ a 和 Allele _ b 代表靶 SNP 的等位基因。

邻近 M13 序列的 C 或者 G 核苷酸能保证不同序列的多重连接产物是平等扩增的。

(3) 以最低合成范围至 100μmol/L 的浓度重悬未修饰的基因分型寡核苷酸链。

(4) 通过每个含 32 个核苷酸的上游寡核苷酸链（100μmol/L）2μL 加水 136μL 配制 1μmol/L 的 16-plex 上游 OLA 寡核苷酸混合物。

(5) 单独对下游寡核苷酸链进行 5′磷酸化。这样做是让寡核苷酸的上游和下游能够连接而又不一同完成（寡核苷酸链之间的相互反应能阻止每个寡核苷酸相等的磷酸化）。

12.5μL 反应体系

| | |
|---|---|
| 水 | 8.125μL |
| T4 多聚核苷酸激酶缓冲液 | 1.25μL |
| ATP（100mmol/L） | 0.125μL |
| T4 多聚核苷酸激酶（10units/μL） | 1μL |
| 下游寡核苷酸（100μmol/L） | 2μL |

置 37℃温育 1h，接着置 65℃温育 20min 结束反应。

(6) 按照 16 个下游寡核苷酸每个混合 12μL 的磷酸化反应物的比例配制 1μmol/L 的 16-plex 下游 OLA 寡核苷酸混合物。

## OLA 反应

(7) 配足量 3μL 反应体系

| | |
|---|---|
| 水 | 2.3μL |
| OLA 缓冲液（10×） | 0.3μL |
| DTT（25 mmol/L） | 0.3μL |
| Taq DNA 连接酶 | 0.04μL |
| 上游寡核苷酸链混合物 | 0.03μL |
| 下游寡核苷酸链混合物 | 0.03μL |

(8) 在 384 孔或 96 孔的 PCR 板上，用重复吸液器或者液体处理器每孔加 3μL 的上述试剂混合物。接着用液体自动处理仪或者手工并行输入工具加入 0.2μL 经 PCR 扩增包含目标 SNP 的 DNA。用并行输入工具可使样本之间均一。一些 PCR 样本应该设不含 DNA 的 PCR 空白对照。

(9) 封板并短暂离心。用下列循环方案完成连接。

i. 预变性 95℃，5min。

ii. 95℃、30s、45℃、25min，3 个循环。

iii. 4℃保存。

## OLA 扩增反应

（10）配足量 12μL 反应体系

| | |
|---|---|
| 水 | 10.196μL |
| 扩增缓冲液（10×） | 1.2μL |
| dNTP（25 mmol/L） | 0.024μL |
| Taq DNA 聚合酶（5 units/μL） | 0.1μL |
| M13F. BRL（50μmol/L） | 0.24μL |
| M13R. BRL（50μmol/L） | 0.24μL |

（11）每个 OLA 连接反应中直接加入 12μL 的试剂混合物，封板，并短暂离心。用下列循环方案扩增连接产物。

i. 预变性 94℃，2min。

ii. 94℃、25s，58℃、35s，72℃、35s，32 个循环。

iii. 72℃、2min。

iv. 4℃保存。

空白（无 DNA 的 PCR）和阳性样本电泳：在阳性样本中可见一条亮带（78bp），在空白对照中则无。因为 M13 引物浓度在 OLA 扩增反应体系中非常高，空白样本通常包含一个副带即引物二聚体带，它比阳性带要小。

## 分析

（12）将扩增反应物置于热循环仪，65℃，晾干，约 1h。

（13）每孔中加入 5μL 经无菌过滤器过滤的变性缓冲液，再按下列循环方案在热循环仪上使样本重悬/变性。

i. 65℃，15min。

ii. 95℃，5min。

（14）用阵列加样器将样本点于尼龙膜上（手工并行输入工具也可以，但是会影响点的质量）。戴无粉实验室手套，不要用未戴手套的手接触膜。让膜上的点晾 10min。以 50mJ 紫外线交连照射样本，并轻摇置于中性缓冲液中的膜，30min（用中和高 pH 的变性上样缓冲液）。

## 杂交

（15）在杂交管中加入膜（多重膜可叠加在单一管中）。加入 5mL 预温至 42℃的杂交缓冲液和经声裂的鲱鱼精子，该精子已经经 96℃温育 5min 后变性。

（16）杂交管在杂交恒温箱中以 4r/min 的速度旋转过夜（首次使用的膜）或者 42℃转 3h（所有接下来的杂交）。

（17）从如下的 10μL 反应中制备用［$\gamma$-$^{33}$P］ATP 末端标记放射标记的寡核苷酸

探针。

| | |
|---|---|
| 水 | 5μL |
| T4 聚核苷酸激酶缓冲液（10×） | 1μL |
| 探针寡核苷酸（10μmol/L） | 1μL |
| T4 聚核苷酸激酶（10 units/μL） | 1μL |
| [$\gamma$-$^{33}$P] ATP（10μCi/μL） | 2μL |

该反应 37℃孵育 40 min，接着 80℃15min 结束反应。用前不必再专门（或另外）纯化该反应。

(18) 向杂交管中加入放射标记探针反应，并以 4r/min 的速度在杂交恒温箱中 42℃转 4h。

### 漂洗

(19) 向废料收集器中排空杂交缓冲液/放射标记探针。用小量预温至 40℃漂洗缓冲液简单地冲洗，并将洗涤液倒入废料收集器。

(20) 加入 50mL 预温（40℃）漂洗缓冲液。在 40℃以 4r/min 的速度旋转管 20min，并将缓冲液倒入废物收集器。循环洗涤 3～5 次。循环洗涤数取决于每管中杂交过滤的杂交量。洗涤次数太少会导致经过膜表面的放射线背景水平不均一。

(21) 从杂交管中移去膜，并用预温至 40℃的漂洗缓冲液冲洗。

### 数据收集

(22) 将杂交/漂洗的膜在荧光屏上暴露 3～4d（实际的时间取决于杂交膜上的点密度）。用荧光图像仪扫描屏。

(23) 用图像采集软件包（如 GE Healthcare Array Vision）分析图像。从免费统计程序包 R(http://www. r-project. org/)获得 custom 软件(http://cstern. bio. uci. edu/tools/genotyping. htm)，进行基因分型调入。

### 洗脱

膜一旦成像，在重新做成探针之前必须洗脱。

(24) 将膜从荧光屏暗盒移至一个中性缓冲液中，以使膜保持湿润。如果反射活性探针连接时膜是干燥的，探针会和膜永久固定。

(25) 将膜置于含有预热至 80℃的 50mL 洗脱缓冲液的杂交管中。在 80℃以 4r/min 的速度旋转管 15min。

(26) 将洗脱缓冲液倒入废料收集器，膜保存在中性缓冲液置 4℃，或者从步骤 (15) 再次开始实验。

## 方案四 用荧光偏振检测法进行模板介导染料掺入分析

Connie Ha[1] and Pui-Yan Kwok[1,2]

[1]*Cardiovascular Research Institute and Center for Human Genetics, University of California, San Francisco, California 94143-0793;*
[2]*Department of Dermatology, University of California, San Francisco, California 94143-0793*

一个分子的荧光偏振是与该分子的转动张弛时间（旋转 68.5°所用的时间）成比例的。转动张弛时间还与溶液黏度、绝对温度和分子的摩尔体积密切相关（Kwok 2002）。如果前两个特性保持不变，那么，荧光偏振就直接与分子的摩尔体积成比例，而摩尔体积又直接与相对分子质量成比例。这样，经适当波长的偏振光激发，大荧光分子在空间里的翻转会相对慢些，并且能观察到偏振发射，而小荧光分子的翻转则快些，并且观察不到偏振发射（消偏振）（图 11-2）。

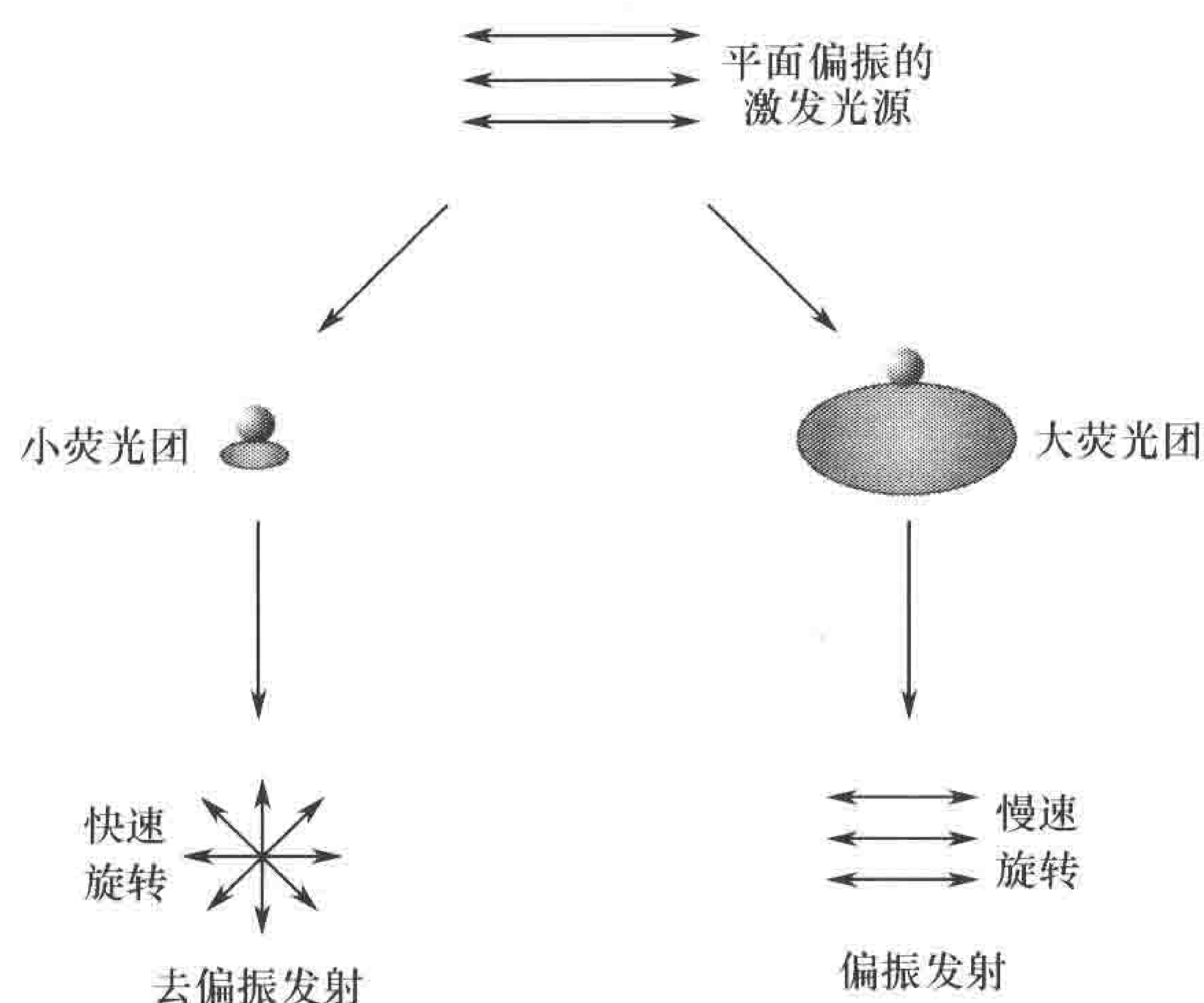



图 11-2 荧光偏振原理描述略图。经适当波长的偏光激发，大荧光分子在空间里的翻转会相对慢些（偏振），而小荧光分子的翻转快些（消偏振）

该实验描述了基于荧光偏振的单个碱基引物延伸的一种变异。用荧光偏振检测法行模板介导染料掺入（FP-TDI）是一种经济、稳健的 SNP 基因分型方法，该法易于优化和实施。这种同质分析用的是未标志、未纯化的 PCR 阶段的引物，应用非常广，既能用于单标记分析的研究，也能用于中通量研究。FP-TDI 分析是一种双脱氧链终止测序 DNA 的实验，该实验对位于靶 DNA 多态位点的上游即刻退火的未标记 SNP 引物一个碱基 3′端进行了即刻辨别。该操作一是利用了 DNA 聚合酶的特异性，该酶能通过染料终止子与靶 DNA 上发现的多态核苷酸互补来延伸退火引物；二是利用了当荧光染料在引物延伸反应中成为较大分子的部分时，FP 的增加。（Chen et al. 1999）FP 是以下面观察为基础的：当被平面偏振光激发后荧光分子会向一个相对于分子自身来说是固定的平面发射出偏振荧光（Perrn 1926）。

TDI 分析由 4 个关键步骤组成，所有步骤可在同一微量滴定板上完成，无须进一步分离和纯化（图 11-3）。首先，基因组 DNA 经 PCR 扩增产生引物延伸反应模板。其次，将含有焦磷酸酶的 PCR 清除多种酶混合液直接加入 PCR 产物中以除去过量的 PCR 引物、脱氧核苷三磷酸和无机焦磷酸。加入焦磷酸酶能使染料终止子的错并入作用最小（Xiao et al. 2004）。再次，用靶 DNA 多态位点上游一个碱基的退火 SNP 引物完成单碱基引物延伸。最后，最终产物用一个 FP 读板器扫描，以决定荧光偏振的变化。

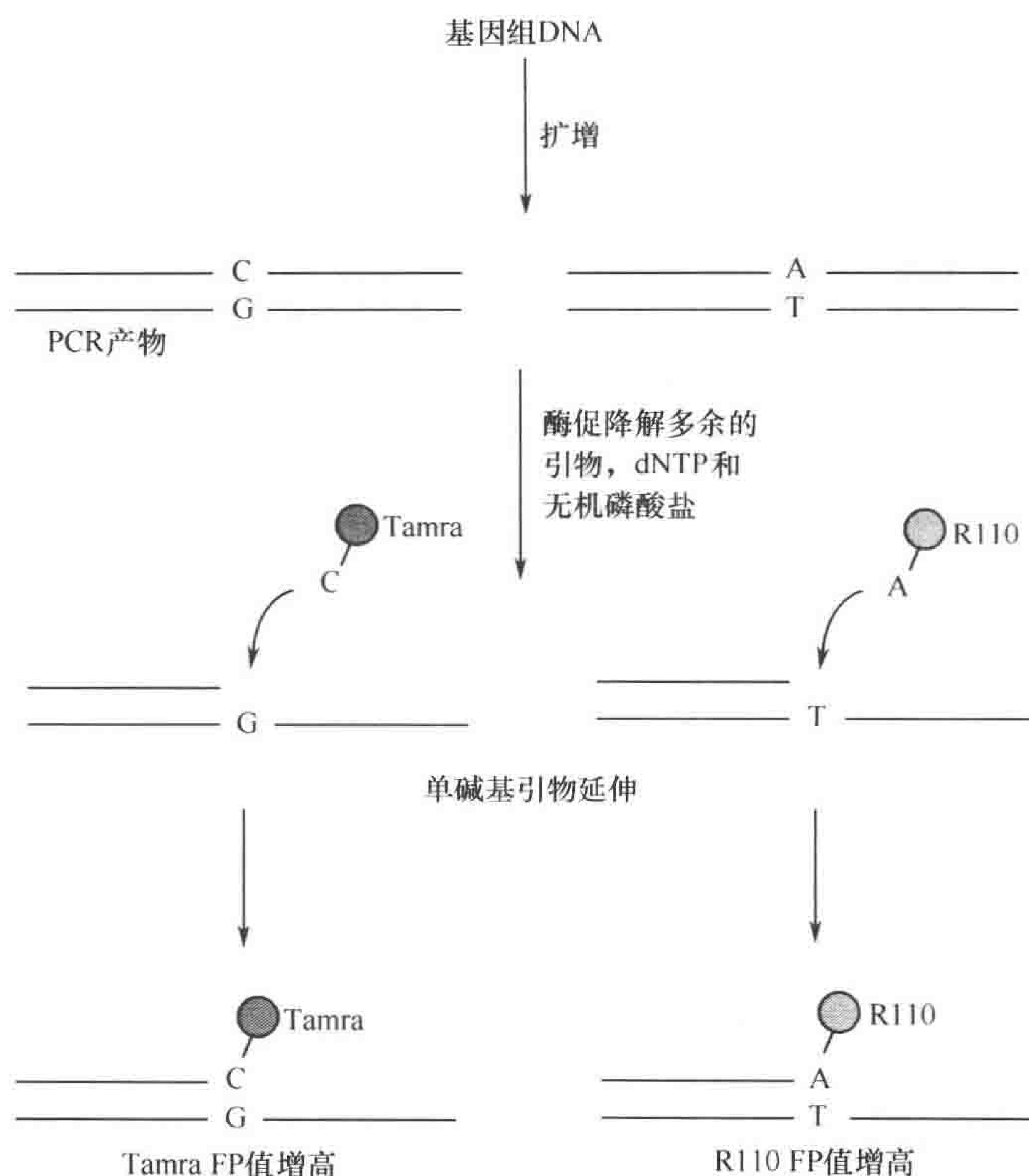



图 11-3　FP-TDI 分析法

PerkinElmer AcycloPrime™ Ⅱ-FP SNP 检测系统是基于 FP-TDI 原理的一种经济的试剂盒。可从 PerkinElmer Life Sciences 得到。试剂盒不采用双脱氧核糖核苷酸，而是包含了 4 种非环三磷酸核糖核苷酸（厂家称为非环终止子），其中的两个用荧光染料标记（图 11-4）。试剂盒还包含了 Acyclopol，一种比双脱氧终止子优先并入非环终止子的耐热 DNA 聚合酶突变体。6 个 AcycloPrime 试剂盒就可以覆盖 6 个可能的等位基因的连接（G/A、C/T、G/T、C/A、A/T 和 G/C）。

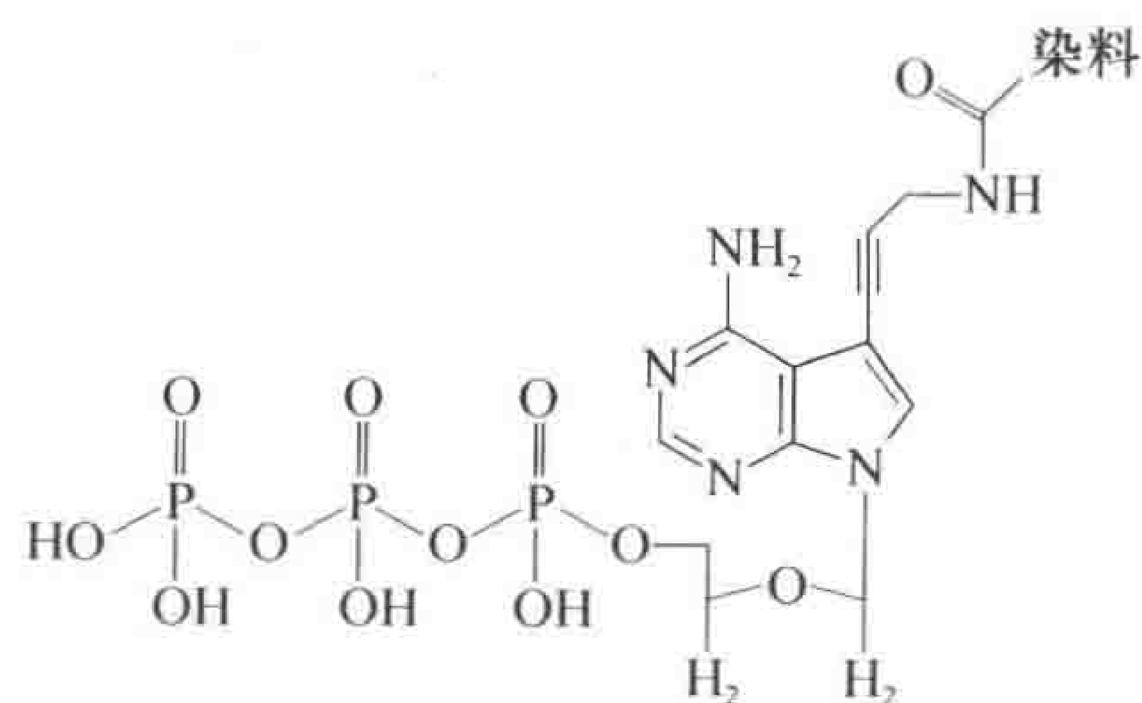



图 11-4　PerkinElmer 公司染料标记的 Acyclo 终止子结构

这种基因分型分析是普遍的、高度特异的且很有效，特别是以分析发展的观点来看。另外，FP 的实施是稳健且容易的；在主要的测序四步骤中的任一步，反应板可以

冷冻和保存，之后仍可用于下次测序。实验中焦磷酸酶的掺入和为分析荧光数据而引进的淬灭使高达95%的分析不用优化就能顺利进行。

FP-TDI 方法是用于 HapMap 计划中的众多基因分型方法之一，并且能产生超过15 000个 SNP 的数据（HapMap 国际性协议 2003）。网上资源的收集包括公共数据库和分析工具，表 11-2 已给出。

**表 11-2 互联网资源**

| 公共的 SNP 研究资源 | |
|---|---|
| dbSNP | http://www.ncbi.nlm.nih.gov/SNP/ |
| SNPper | http://bio.chip.org:8080/bio |
| SNP consortium | http://snp.cshl.org/ |
| PerkinElmer | http://perkinelmer.com/ |
| 免费的引物设计软件 | |
| RepeatMasker | http://ftp.genome.washington.edu/RM/ |
| Primer3 | http://www.genome.wi.mit.edu/genome_software/ |
| 数据输出 Excel 宏 | |
| PerkinElmer | www.snpscoring.com |

# 材料

注意：标有〈!〉标志的材料的适当处理，见附件。

## 试剂

AcycloPrime™-FP SNP 检测系统（PerkinElmer）包括下列试剂

- 10×反应缓冲液
- PCR Clean-Up 试剂
- PCR Clean-Up 缓冲液（10×）
- Acyclopol 酶，用作单碱基延伸
- Acyclo 终止子混合物，包括等量的染料标记的 R110 和 Tamara 终止子
- 扩增缓冲液（10×）

适合于分子生物学的双脱氧核糖核苷酸仪器（dNTP，每种是100mmol/L；Sigma-Aldrich）

$MgCl_2$

寡核苷酸链，SNP 序列和引物设计软件

PCR 引物寡核苷酸链（从 Intergrated DNA Technology [IDT]，Invitrogen，或 Sigma-Aldrich 获得）

SNP 引物通常是18～25多聚寡核苷酸链，与 PCR 引物的正义链或是反义链互补，设计成靶 DNA 多态位点上游3′端立即退火（SNP 引物可从 Intergrated DNA Technology 购买）。

用作引物设计的 SNP 序列从公共数据库，如 dbSNP（http://www.ncbi.nlm.

nih. gov/ SNP/), SNPper (http://bio. chip. org:8080/bio/)或 SNP consortium(TSC: http://snp. cshl. org/)获得。另外，要分析超过 38 000 的引物设计的话，可免费从 PerkinElmer 数据库获得(http://perkinelmer. com/)。

引物设计可用免费获得的软件，如 RepeatMasker (http://ftp. genome. washington. edu/RM/)和 Primer3 (http://www. genome. wi. mit. edu/genome _ software/)。

焦磷酸酶 (1mg, Roche)

Taq DNA 聚合酶 (Platinum Taq) (5 units/μL 用 10×扩增缓冲液; Invitrogen)

### 器材

离心机。各种离心机（带有板适配器，每加入一种试剂后可用它压板)，附件可从 Fisher 或 VWR 购买。

多重标记读板器 (EnVision™多重标记读板器，PerkinElmer；或者 Victor 多重标记器，PerkinElmer)，在 TDI 步 [步骤 (10) 和步骤 (11)] 之后读板器通过读取发射的荧光产生数据文件。

PCR 板，黑色 384-或 96 孔板 (LabSource or USA Scientific)。不要用光亮的板。PCR 板必须是黑色，以便多重标记读板器能够检测出 FP 的发射，PCR 板必须与热循环仪和多重标记读板器相适应 。

PCR 封闭包，384 孔 (LabSource)。封闭包如果用 5%的漂白粉冲洗并完全漂洗干净可再利用。偶尔用新的封闭包代替，取决于使用量。封闭包可用 PCR 薄膜或箔黏合剂代替 (Thermo Fisher Scientific 或 VWR International)。

PCR 热循环仪（如 Auto-Lid Dual 384-Well GeneAmp PCR System 9700 或 96-Well GeneAmp PCR System 9700, Applied Biosystems)。对核苷酸扩增，有许多不同的热循环选项设计，可从 Fisher 或 VWR 获得。

移液台 (Evolution P³ Precision Pipetting Platform, PerkinElmer), Evolution P³是分配试剂的快速通量的液体自动处理系统。依据试验设计，可用常规单道或者多重通道的移液器实施。

## 方法

### 引物设计

(1) 通过向 dbSNP 或者其他数据库递交 reference SNP (rs) 或 submitted SNP (ss) 序列号获得靶 SNP 周围的序列。RefSNP accession ID (rs) 和 submitted SNP record (ss) 序列号是 dbSNP 对向公共数据库递交的 SNP 的命名。在 dbSNP 网址上依照步骤递交 rs 和 ss 序列号可轻松获得 SNP 的侧翼序列。

(2) 用软件 (RepeatMasker) 分析获得的序列，要避免在重复区域选择引物 (Vieux et al. 2002)。该软件有详细的注释且易于使用。用唯一的基因组序列挑选 PCR 引物能提高 PCR 序列设计的成功率。

(3) 在软件 Primer3 中用一组最佳参数选择引物 (表 11-3)。

表 11-3 引物设计的 Primer3 参数

| Primer3 参数 | |
|---|---|
| 引物产物片断范围/bp | 80～400 |
| 引物产物最优大小/bp | 250 |
| 引物大小/bp | 最小：20 最优：23 最大：26 |
| 引物 $T_m$/℃ | 最小：54 最优：55 最大：56 |
| 引物 GC/% | 最小：20 最大：50 |
| GC Clamp | 0 |
| DNA 浓度/（mmol/L） | 40 |
| 引物盐浓度/（mmol/L） | 50 |
| 引物对 wt 产物大小 LT（部分） | 0.20 |
| 引物对 wt 产物大小 GT（部分） | 0.50 |
| 引物 self any | 8 引物 self end： 3 |
| 引物最大末端稳定度 | 8 |
| 引物 explain flag | 1 |
| 引物 num return（引物对） | 1 |
| 目标起始位置/bp | *X* |
| 目标长度/bp | 50 |

## 扩增反应

要使 FP-TDI 分析进行得很好，必须产生有效的 PCR 产物以供引物在延伸反应中使用，必须使剩余 PCR 引物和 dNTP 量最小以使用作 PCR 清除步骤的酶有足够能力降解多余的引物和 dNTP。因此设计该实验时要考虑到这些要求。

（4）在 96 或 384 孔的反应板中每孔加入 2.4ng 的基因组 DNA；空气中过夜，晾干。用薄纸覆盖板以免灰尘落入。一般来说，我们推荐每个 96 孔的组留两个空孔，且包含 4 组重复产物作为对照。干燥的 DNA 非常稳定，能保存一年，这取决于原血样的质量。用塑料封皮包裹干 DNA，储存在干燥器中。

（5）制备 PCR 酶混合液

| | |
|---|---|
| 10×扩增缓冲液 | 0.5μL |
| $MgCl_2$ 50mmol/L | 0.25μL |
| dNTP 混合物，每种 2.5 mmol/L | 0.1μL |
| Platinum Taq DNA 聚合酶（5 units/μL） | 0.02μL |
| 水 | 2.13μL |

加酶时要迅速，在步骤（6）分装该混合物之前，所有溶液在冰上操作。

（6）将 3μL PCR 引物混合物（每个上游和下游引物，各 0.2μmol/L 和 3μL PCR 酶混合液混加起来。将此总体积为 6μL 的 PCR 混合物分配到装有干 DNA 的每个孔中［步骤（4）中的板］。为使污染减小到最小，将此混合液加在每孔的上部，离心，封孔。

如果不只用一个标记进行基因分型，最好分别加 3μL 引物混合物和 3μL 的 PCR 酶混合液。每次加完一种液体后离心，使污染减小到最小。封板避免样本从板边缘蒸发。每板仔细做好标记，因为同样的板要用来完成整个实验。

PCR 混合物相对稳定，能作后续使用。如果不被反复冻融，过量 PCR 混合物可在 −20℃储存 3～4 周而不降解。

(7) 预热热循环仪热盖到 105℃。依下列程序完成扩增。

| 循环数 | 变性 | 退火 | 聚合/延伸 |
|---|---|---|---|
| 1 | 95℃，2min | | |
| 45 | 92℃，10s | 58℃，20s | 68℃，30s |
| 最后 | | | 68℃，10min |

储存于 4℃或者直接进行步骤（8）。

有必要将热循环仪的顶盖预热，并使其保持受热，这样当整个程序运行时，能阻止循环中蒸发。如果热循环仪的顶盖温度比板的温度低，蒸气冷凝会污染整个板。

## 含焦磷酸酶的 PCR 清除

PCR 产物必须经过降低反应中产生的焦磷酸、过量的 PCR 引物和 dNTP 的处理。这是必要的，因为引物、dNTP 和焦磷酸酶会干扰引物延伸反应（Xiao et al. 2004）。通过添加三种酶来完成降解：核酸外切酶Ⅰ、虾碱性磷酸酶和焦磷酸。核酸外切酶Ⅰ、虾碱性磷酸酶作为组合试剂是由厂家提供的。

(8) 以 10.5∶1.33∶1.5 的比例组合 10×PCR 清除缓冲液、PCR 清除试剂和焦磷酸酶，制备 PCR 清除混合物。吸取 2μL 的清除混合物加于步骤（7）中的每个 PCR 产物中。离心，封孔。如果不反复冻融的话，PCR 清除混合物可以在−20℃储存 2 周。

(9) 预热热循环仪盖子到 105℃。于 37℃温育上述反应混合物 1h 并通过 90℃加热 15min 使酶失活。将该反应体系储存于 4℃。

## 引物延伸（TDI）

引物延伸反应是基于一个碱基的 DNA 测序。因此该实验是 DNA 循环测序的实验。

(10) 用下列成分制备 TDI 的酶混合液以备引物延伸反应使用。

| | |
|---|---|
| Acycloprime10×反应缓冲液 | 2 μL |
| Chosen 染料终止子混合物 | 1 μL |
| Acyclo 酶 | 0.05μL |
| 水 | 4.95μL |

(11) 将 8μL 的 TDI 酶混合液和 5μL 的 SNP 引物（1μmol/L，上游或者下游 SNP 引物）混合物进行组合，向每个后-清除 PCR 产物中加入此总混合物。离心，封板。

如同在步骤（6），如果对多个分子标记同时进行基因分型，最好分别加 5μL 的 SNP 引物混合物和 8μL 的 TDI 酶混合液。在每次加完一种液体后离心，以使污染减小

到最小。封板避免样本从板边缘蒸发。

如果不被反复冻融，过量的 TDI 酶混合液可在－20℃储存 3～4 周。TDI 酶混合液对光敏感，储存前用箔纸仓裹。

（12）预热热循环仪顶盖到 105℃。依下列程序完成延伸反应。

| 循环数 | 变性 | 聚合/延伸 |
| --- | --- | --- |
| 1 | 95℃，2min | |
| 5～15 | 95℃，15s | 55℃，30s |

储存于 4℃。

### 读板和数据分析

（13）将板从热循环仪上移出，离心。揭去封板膜，将板置于 Envision 读板器上。

（14）从 Envision 读板器输出 FP 文件，用 excel macro EnVisionMacro _ Excel384 _ 4×96 运行，产生 excel 格式的文件。

（15）依据数据集和 FP 值进行基因分型调入，好的数据应该在散点图上显示 4 个明显的集簇。阴性对照（孔里无 DNA）的 R110 和 TAMRA 值低并且靠近原点，表明染料终止子在溶液中是单体的。纯合子的数据点应该是 FP 值在 R110 高而在 TAMRA 低，而杂合子的数据点是 FP 值在 R110 和 TAMRA 中都高。

（16）对于数据中 R110 出现激烈衰减的情况，通过设计 TAMRA FP 值对 R110 密度比例重新分析荧光数据，可能增加基因分型的调入（图 11-5）。TAMRA 值对 R110 密度比例（R）定义为

$$R = 100 \times \mathrm{I}_{\mathrm{TAMRA总值}} / \mathrm{I}_{R110\mathrm{总值}}$$ (Xiao et al. 2004)

（17）如果散点图显示 4 个集簇分离不完全且 FP 值在 R110 和 TAMRA 都低，将板重新返回热循环仪进行更多次 TDI 循环 [如同步骤（12）]，完成反应可能对此有所帮助。

## 疑难解答

问题（步骤（10）、（11），引物延伸反应）：用 FP-TDI 进行基因分型是一种稳健、精确的分析方法，但是观察到一些由用于本实验的染料终止子的特性带来的固有问题。分析失败的一类原因是 R110 标记的非环终止子低 FP 信号导致了集簇的分离差。

解决方法：在一些延伸产物中出现的低 R110 FP 值可用 R110 掺入 SNP 引物后淬灭来解释。对这种现象的系统调查揭示了荧光淬灭和因不同染料偏振产生的差别之间的关系。荧光淬灭大部分只发生在 R110 标记的非环终止子中，某些分析中，荧光强度遭到淬灭可多达 90%（Xiao et al. 2004）。在我们的研究中，所有 TAMRA-非环终止子的荧光密度依其掺入 SNP 引物增多而增加，且非常稳定，FP 值超过 120mP。相反，R110-非环终止子的荧光密度依其并入 SNP 引物增多显著地降低，同时显示了 FP 值变

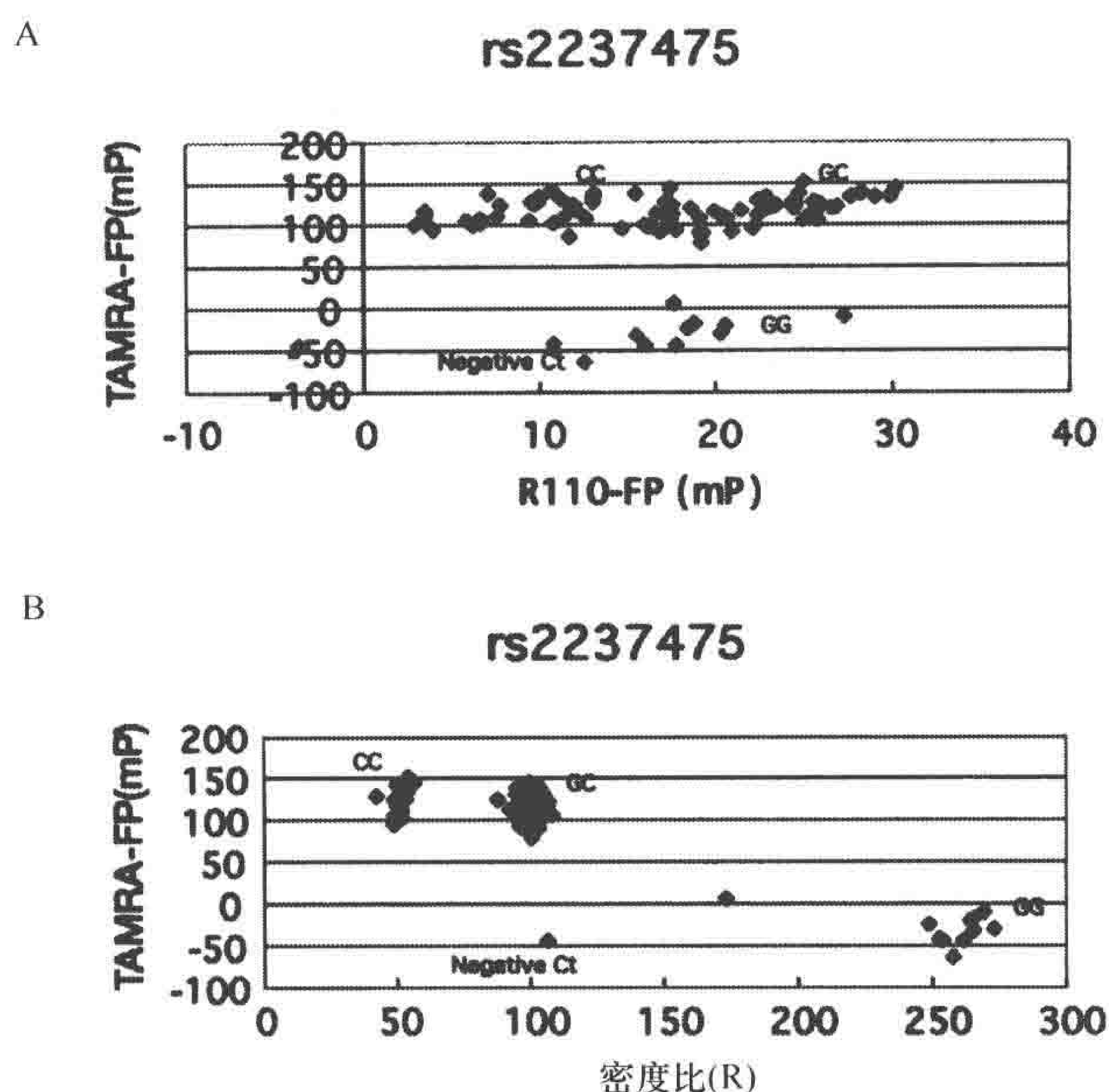



图 11-5　通过设计 TAMRA FP 值对 R110 的密度比例对分子标记 rs2237475 进行淬灭分析。A. 低的 R110 FP 值导致差的集簇分布；B. 同一组数据用 TAMR AFP 值对 R110 密度比例（R）进行重新设计

化范围 30～150mP。淬灭效应在那些 3′端是鸟嘌呤或者距 3′端 10 个碱基内有若干鸟嘌呤的序列中能进一步加剧（Xiao and Kwok 2003）。对 R110 严重淬灭的数据分析将导致 FP 值的丢失，且这种分析还会成为一种染料-终止子掺入的差的显示计。经 Acyclo-pol 相对低的 R110 非环终止子的掺入引起 R110 标记问题（Xiao et al． 2004）。确认并考虑淬灭现象后我们能容易地恢复荧光数据，调入好的基因分型。正如在早些出版物中详尽论述的一样（Xiao et al. 2004），检测真正的 R110-非环终止子掺入的一个好方法是设计 R110 淬灭的程度。通过设计 TAMRA FP 值对 R110 密度比例（淬灭），而非 R110 FP 值，重叠的组会分离为 4 个清晰的集簇。

问题（步骤（10）、(11)，染料标记终止子的错掺入)：在 FP-TDI 实验中提出的另一个警告是引物延伸反应中，染料标记终止子之一的错掺入会导致分析失败。属于这类的分析失败的标志是一个纯合簇的缺失伴随不寻常高数量的杂合子。我们注意到当在反应中 SNP 引物与非环终止子以同样的碱基结束时，基因分型轮廓会落入前面描述过的类型中。

解决方法：在琢磨失败原因、改进分析的过程中，我们发现隐藏在染料-终止子错掺入背后的罪魁祸首是在 PCR 过程中产生的过多无机焦磷酸（PPi）。在 PCR 时产生的大量 PPi 能导致焦磷酸解作用，这是与 DNA 聚合酶催化 SNP 引物 3′端碱基裂解相反的反应。经这种机制缩短的引物再经染料-终止子的延伸与靶序列互补，从而产生了错误的基因分型。我们的观察表明由焦磷酸导致的错掺入经常发生在一种终止子被用完而前面的延伸反应完成之后（Xiao et al. 2004）。引物延伸反应之前降解 PPi 可以部分地

预防焦磷酸解作用，因此也是成功分析的一部分。在研究中，我们通过在 PCR 清除步骤中用焦磷酸酶的混合物孵育 PCR 产物除去 PCR 过程中产生的 PPi，已经能够有效预防非环终止子错掺入（图 11-6）。当反应中 SNP 引物不是以与染料-终止子终止末端相同的碱基结束时，未标志的终止子错掺入引物种并不会干扰反应中的荧光偏振。

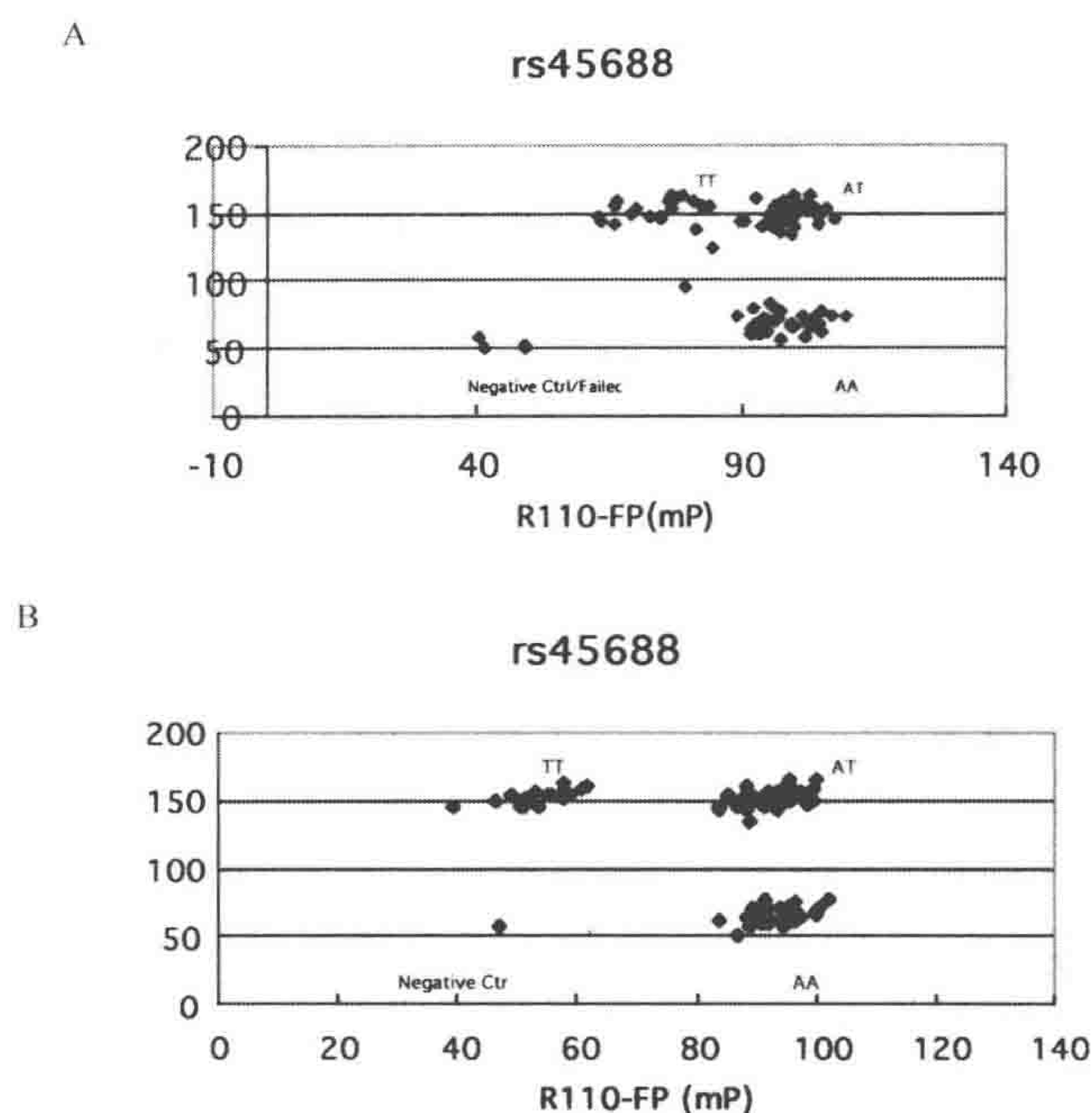



图 11-6 在 PCR 清除步骤的温育一步未加入分子标记 rs45688 及焦磷酸酶的基因分型数据。A. 在 30TDI 循环中没有焦磷酸酶孵育的基因分型结果。纯合子 TT 簇移向杂合子 AT 簇并开始与其融合。B. 同样的反应有焦磷酸酶温育的基因分型结果。甚至在达 70 个循环后依然观察不到错掺入

## 致谢

我们感谢 Angie Phong 的图和 Xiao Ming、Chan Ting-Fung 博士们于本文的意见。

## 方案五 温度梯度毛细管电泳分析

An-Ping Hsia,[1] Hsin D. Chen,[1] and Patrick S. Schnable[1,2,3]

[1]*Department of Agronomy,* [2]*Department of Genetics, Development, and Cell Biology,* [3]*Center for Plant Genomics, Iowa State University, Ames, Iowa 50011*

本方案描述用温度梯度毛细管电泳（TGCE）检测两个 DNA 片段之间的多态性（Hsia et al. 2005）。在个体内或者个体之间检测多态性，取决于实验设计。在这种分析中，目的等位基因和相关等位基因都经 PCR 扩增，高保真 Taq 聚合酶与随之产生的扩增子以 1∶1 比例混合。这种混合物是经变性和退火的，以形成同源双链（在目的等位基因和相关等位基因之间不存在多态位点）和杂交双链（在感兴趣等位基因和相关等

位基因之间存在多态位点）。然后用能同时分析 4 个 96 孔的 Reveal System（SpectruMedix）进行电泳和检测，并用 Revelation 软件储存，看数据。

## 材料

### 试剂

AmpliTaq Gold（5 U/μL，Applied Biosystems）
克隆的 Pfu DNA 聚合酶（2.5 U/μL，Stratagene）
DNA 样本
dNTP（2mmol/L，Intermountain Scientific）
上游引物（5μmol/L）
$MgCl_2$（25 mmol/L）
矿物油
PCR 缓冲液（10×）
下游引物（5μmol/L）
TGCE 试剂（SpectruMedix）
  瓶 1 和 2：毛细管洗液（WASH-500-002）
  瓶 3：Reveal 无染料的电泳缓冲液（BRUR-500-002）
  瓶 4：Reveal Mutation Discovery 电泳缓冲液（BRUR-500-001）
  胶：Matrix，Reveal Mutation Discovery（MREV-240-001）

### 器材

Reveal System，model RVL 9612，rev. 2.0（SpectruMedix）

Revelation 分析软件（2.4 版）显影、记录 TGCE 数据（SpectruMedix）。现在可通过 Transgenomic，Omaha，Nebraska 得到 Reveal System 和 Revelation 软件。

GRAMA 软件（Maher et al. 2006）

## 方法

（1）为靶等位基因和相关等位基因制备 PCR 反应混合物（总体积 20μL）。依据实验设计，可选同一式不同基因组之内的靶等位基因和相关等位基因。

| | |
|---|---|
| 20～50ng 基因组 DNA | 2.5μL |
| 10×PCR 缓冲液 | 2.0μL |
| 2 mmol/L dNTPs | 2.0μL |
| 25 mmol/L $MgCl_2$ | 1.6μL |
| 5μmol/L 上游引物 | 2.0μL |
| 5μmol/L 下游引物 | 2.0μL |
| 5 U/μL AmpiTaq Gold | 0.09μL |

2.5 U/μL Pfu 聚合酶　　　　0.02μL

水　　　　　　　　　　　　7.79μL

(2) 用下列的 PCR 程序扩增靶等位基因和相关等位基因。

i. 95℃，10min。

ii. 94℃，3min。

iii. 94℃，30s。

iv. 60℃，45s。

v. 72℃，1min30s。

vi. 返回到步骤 iii，30 次。

vii. 72℃，10min。

viii. 维持 12℃不变。

(3) 取靶等位基因和相关等位基因的 PCR 反应物 5μL。在热循环仪中用下列程序使样本变性并退火。

i. 95℃，2min 40s。

ii. 95℃，20s（重复 15 次，每次温度降低 1℃）。

iii. 80℃，1min（重复 25 次，每次温度降低 1℃）。

iv. 55℃，18min。

v. 55℃，1min（重复 10 次，每次温度降低 1℃）。

vi. 45℃，30s（重复 10 次，每次温度降低 1℃）。

vii. 维持 12℃不变。

(4) 向 PCR 混合物中添加 10μL 的矿物油以阻止样本在电泳时挥发。装载板并依下列 TGCE 说明步骤运行。

i. 瓶＃1，线清除，清除率：20.0 mL/min，15 min。

ii. 瓶＃1，流速：22.5 mL/min，5 min。

iii. 瓶＃2，流速：25.0 mL/min，5 min。

iv. 瓶＃3，线清除，清除率：10.0 mL/min，3 min。

v. 瓶＃3，流速：6.0mL/min，6min。

vi. 凝胶，线清除，清除率：12 mL/min，12 min。

vii. 凝胶，胶注射，Volume delivered/cap：42μL。

viii. Delivery 时间：10 min。

ix. 瓶＃4，凝胶预试，流速：5.0 mL/min，10.0 kV，5 min。

x. 瓶＃4，样本注射，流速：5.0 mL/min，6.0 kV，50s。

xi. 瓶＃4，电泳＋电流监测器，流速：5.0 mL/min，9.0 kV，15min。

xii. 瓶＃4，数据采集，流速：5.0 mL/min，9.0 kV，50 min。

在这步进行测量和作图时分别用 50 min 和 30 min。

(5) 用 Revelation 软件包［理想的话，用 GRAMA 软件（Maher et al. 2006）做补充］记录和观察数据。软件的详细描述和数据输出展示及分析举例见第 26 章。

## 致谢

该项目得到国家科学基金植物基因组计划竞争基金资助（OBI-0321711）和 Hatch Act 基金及 IowaⅢ基金资助。

## 方案六　源于玉米 454 EST 序列的 SNP 发掘

W. Brad Barbazuk,[1] Scott Emrich,[2,3] and Patrick S. Schnable[4,5,6]

[1]*Donald Danforth Plant Science Center, St. Louis, Missouri 63132;* [2]*Bioinformatics and Computational Biology Graduate Program,* [3]*Department of Electrical and Computer Engineering,* [4]*Department of Agronomy,* [5]*Department of Genetics, Development, and Cell Biology,* [6]*Center for Plant Genomics, Iowa State University, Ames, Iowa 50011*

经 454 Life Science 公司商业化的大量类似焦磷酸测序的技术可用来对玉米 B73 茎干（尖）顶端分生组织（SAM）的转录进行测序 isolated 用激光捕获显微切割法（LCM）（指茎尖的获取方法）。数据分析显示：LCM 与浓缩 454 技术的深测序方法的结合更加丰富了 SAM 转录产物的分析，此转录产物在当前 EST 收集物中尚未出现过。RT-PCR 曾用来验证 27 个基因的表达，这 27 个基因的表达已经通过 LCM-454 技术在 SAM 中检测到，但是在 GenBank 中缺乏直系同源。显著地，这些经验证 SAM-表达"孤儿"中 74%（20/27）能在 SAM 中检测到，而在富含分生组织的不成熟耳中却检测不到。我们得出这样的结论"LCM 和 454 测序技术联合有利于发现稀少"，很可能是源于特殊细胞类型的转录。

在本实验中，454 表达序列标签（EST）是从 454 Lief Science GS-20 测序系统玉米杂交系 SAM 的 cDNA 测序中产生的。接着用计算工具 POLYBAYES 鉴别 SNP。POLYBAYES 已在包括玉米的多个系统中成功地鉴别 SNP（Useche et al. 2001），特别推荐将其用于 454 测序中鉴别 SNPs。关于应用这个计算工具的详细讨论和 POLYBAYES 介导的 SNP 发掘的事例，详见第 26 章。

# 材料

## 试剂

玉米杂交系 B73 和 Mo17 茎干顶端分生组织（SAM）的 cDNA。6～10 SAM（包含 15 000～18 000 个细胞）将提供 10ng RNA。RNA 样本再经两轮扩增，一般能产生 20～60μg 的扩增 RNA（aRNA），按第 8 章描述的制备好。20μg 的 aRNA 约产生 15μg 的 cDNA。

## 器材

BLAST

CROSS_MATCH

454 Life Science GS-20 测序系统（454 Life Science，http://www.454.com）。454 Life Science 有测序服务中心，该中心能提供像 cDNA 和基因组 DNA 这样的样本。关于 cDNA 的质量和数量要求可用它们检查。

POLYBAYES(http://bioinformatics. bc. edu. /marthlab/polybayes. html)

## 方法

(1) 通过对 454 Life Science GS-20 测序系统中的玉米杂交系 B73 和 Mo17 SAM 的 cDNA 进行测序产生 454 EST。

(2) 通过鉴别每一个 454 EST 和基因组序列集之间的最高-评分调准，用 BLAST 向玉米基因组锚定序列分配 454 EST（1e-8 最小 E-值）。

分配好的 EST 可代替基因组 DNA 作为一个锚定序列。主要要求是锚定序列质量要高，因为它们要操纵多重序列调准（MSA）。尽管 EST 将“最佳打击”标准用于分配固定，但是无论是在经 CROSS _ MATCH 形成 MSA 时（见下述），还是在 POLYBAYES 应用内部旁系同源过滤器时都会引起差的分配或者旁系同源之间的分配。玉米 B73 的基因组序列现在已经测序完，这将提供极好的锚定序列集。

(3) 在每一个锚定序列及其相关的 454 EST 上运行 CROSS _ MATCH，以产生一个锚定的 MSA。建议使用下面的 CROSS _ MATCH 参数：

-discrep _ lists-tags-masklevel 5-gap _ init-1-gap _ ext-1。

低启动作用（-gap _ init）和扩大延伸（-gap _ ext）是用来增加短 454 EST 和基因组锚定之间的分配兼容性。如果锚定的 MSAs 是非拼接的，gap _ init 和 gap _ ext 取高值（如 EST 与一个 EST 锚定序列是对齐的，或者基因组序列与基因组锚定序列是对齐的）。

(4) 在 MSA 上运行 POLYBAYES。建议玉米测序使用下列的 POLYBAYES 参数

-maskAmbiguousMatches
-nofilterParalogs
-priorParalog0. 03
-thresholdNative0. 75
-screenSnps
-considerAnchor
-noconsider TemplateConsensus
-prescreenSnps
-priorPoly0. 01
-thresholdSnp0. 5

包含锚定序列和与其对齐的序列（号码序列）是必要的。如果这些序列质量文件是不能够得到或不可信的，则设置错误质量文件

-anchorBaseQualityDefault
-memberBaseQualityDefault

因为在形成 MSA 时 CROSS _ MATCH 将每个序列与锚定序列逐一对齐了，而且 POLYBAYES 能在单个基础上评价碱基质量。使用严格的默认值而非由 454 Life Science 提供的碱基质量信息，期望能增加多态性检测的准确性。

(5) 通过阅读 POLYBAYES 输出文件并决定合适规则以从假阳性中区别推定的 SNP，完成后处理。在由本章作者指导的玉米序列发掘实验中，Mo17 和 B73 454 ESTs 都是可以得到的，并且 B73 玉米 MAGI 组装序列可作为校准的锚定序列。因为 Mo17 和 B73 是同系繁殖的，在每个碱基的位置它们都应该是单等位基因，相对的极个别例外则是由几乎确定的旁系同源（NIP）引起的。因此，用制定的规则过滤假定的 SNP，以大量地减少假阳性率。这些规则是：

i. 多态位点要求在 Mo17-454-ESTs 中最小为 2 倍代表性。

ii. 所有 Mo17 碱基调入的位点，也是在 Mo17 454 EST 和 B73 MAGI 锚定序列之间具有多态性的，是能被预期确定的。这保证了 Mo17-454-EST 的单等位基因性。

iii. 当 B73-454-ESTs 序列经过规则 5 i 和 5 ii 也交叉排列入多态位点时，所有 B73-454-EST 和 MAGI3. 1 锚定序列的碱基调入必须一致。这避免了不正确的 MAGI 碱基调入或者 B73 内的 NIP 造成的多态性。

## 致谢

本项方案的发展受到美国国家科学基金会（基金代号 DBI-0321595，CNS-0521568）、ISV's 植物科学研究所、唐纳德丹福斯植物科学中心的支持。爱荷华州立基金和 Hatch Act 也给予了一定的支持。

## 参考文献

Bonfield J.K., Rada C., and Staden R. 1998. Automated detection of point mutations using fluorescent sequence trace subtraction. *Nucleic Acids Res.* **26:** 3404–3409.

Chen X., Levine L., and Kwok P.Y. 1999. Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res.* **9:** 492–498.

Hsia A.P., Wen T.J., Chen H.D., Liu Z., Yandeau-Nelson M.D., Wei Y., Guo L., and Schnable P.S. 2005. Temperature gradient capillary electrophoresis (TGCE)—A tool for the high-throughput discovery and mapping of SNPs and IDPs. *Theor. Appl. Genet.* **111:** 218–225.

International Hapmap Consortium. 2003. The International Hapmap Project. *Nature* **426:** 789–796.

Kwok P.Y. 2002. SNP Genotyping with fluorescence polarization detection. *Hum. Mutat.* **19:** 315–323.

Maher P.M., Chou H.-H., Hahn E., Wen T.-J., and Schnable P.S. 2006. GRAMA: A genetic mapping tool for the analysis of temperature gradient capillary electrophoresis (TGCE) data. *Theor. Appl. Genet.* **113:** 156–162.

Perrin F. 1926. Polarization de la lumiere de fluorescence. Vie moyenne de molecules dans l'etat excite. *J. Phys. Radium* **7:** 390–401.

Stephens M., Sloan J.S., Robertson P.D., Scheet P., and Nickerson D.A. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.* **38:** 375–381.

Useche F.J., Gao G., Harafey M., and Rafalski A. 2001. High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Inform.* **12:** 194–203.

Vieux E.F., Kowk P.Y., and Miller R.D. 2002. Primer design for PCR and sequencing in high-throughput analysis of SNPs. *BioTechniques* **32:** S28–S32.

Xiao M. and Kwok P.Y. 2003. DNA analysis by fluorescence quenching detection. *Genome Res.* **13:** 932–939.

Xiao M., Phong A., Lum K., Greene R.A., Buzby P.R., and Kwok P.Y. 2004. Role of excess inorganic pyrophosphate in primer-extension genotyping assays. *Genome Res.* **14:** 1749–1755.

# 12　倒置分子探针和基因芯片：应用于高密度标签 SNP 分型

George Karlin-Neumann, Marina Sedova, Ronald Sapolsky, Jonathan Forman, Yuker Wang, Martin Moorhead, and Malek Faham

*Affymetrix, South San Francisco, California 94080*



## 简介

靶向基因分型意味着我们具有可以对基因组中几乎所有感兴趣的位点进行分型的能力。这对于那些科学假说所主导的研究尤其重要，或者应用于确定基因组中特定位置的基因型，如关联研究中的候选基因或区域；或用于确定我们已知或猜测具有决定功能和（或）表型的多态性位点的基因型。这些应用的要求可以分为：低密度，如诊断学（每个样本中含数十、数百到数千个分子标记）（Van Eerdewegh et al. 2002）；中等密度，候选基因研究或关联研究（每个样本中含数千到上万个分子标记）(Zheng et al. 2006)；高密度，功能性多态研究或全基因组关联性研究随访（每个样本中包含数万个分子标记）(Begovich et al. 2004；Shiffman et al. 2005)。在非常高密度的水平，靶向基因分型可以应用于全基因组关联性研究。这依赖于基因组中高的连锁不平衡（LD），仅仅使用一套“标签”SNP 分型来间接地评估大量 SNP（每个样本中包含数十万个分子标记）(Risch and Merikangas 1996；de Bakker et al. 2005；International HapMap Consortium 2005)。

一种适合这些要求的方法必须具备下列特征

- 高转换率和最低限度的探针设计约束条件（如针对大多数标靶序列的探针可

以被设计并顺利完成）。

- 高度的靶特异性。
- 高敏感性。
- 高精确度。
- 高密度水平。
- 高通量。

其他可取的特征可能包括

- 灵活定制化（包括补充标准品）。
- 分型标记的精确定量。

倒置分子探针（MIP）基因型检测技术显示出了所有的这些特点：当使用它验证HapMap的SNP时，首轮转化率高达95%。而且，它可以从量少至1$\mu$g的未扩增人类全基因组DNA样本或50ng全基因组扩增DNA样本中，区别出小到一个碱基不同的两个相似的靶序列。通过与三体家系（父-母-子）一致性或与HapMap基因型的一致性计算，该检测方法通常可以达到99.5%以上的精确度；每个反应的检测密度可以＜1500或＞50 000，杂交之后，可以在10min内扫描在一张Affymetrix基因芯片上，由2或3个人可以在2d内处理完48～96个样本，或者每周192～384个样本。不管是标准品还是定制品都可以制造，或者定制目标SNP增加到标准品中。最终，改良版的芯片可以确定基因型和＜100ng基因组DNA中等位基因的精确拷贝数（第17章）。一种基因分型检测的方法如图12-1A所示，展示了探针退火到样本DNA通过甲基化结合到基因芯片上的过程。

通过结合可环化探针的属性（Nilsson et al. 1994），MIP方法具备了这些特征，那些特征加强酶对于SNP位点的识别，并将其转化为对于一个高度特异的序列片段的识别，然后通过甲基化结合到基因芯片上（Hardenbol et al. 2005）。由于与目标SNP位点序列的双向识别作用（图12-1B），MIP探针达到了特异性的目标，同时它的敏感性和密度分布的水平也很高。退火时，探针两端的同源臂结合到预先已经变性的基因组DNA上，在SNP位点上留下一个碱基的空白供研究。退火反应可以分为4个相同的部分，每个里面加入一种dNTP（第一个里面加入dATP，第二个里面加入dCTP，第三个里面加入dGTP，第四个里面加入dTTP）。每个样品中的等位基因位点决定于给定的探针是否在这种特定碱基存在的情况下发生了环化。在这一步，通过在空白处增加额外的碱基，聚合酶使其特异结合，连接酶使环化的探针闭合，特异性被进一步提高。在随后的核酸外切酶消化阶段，增加了信噪比（S/N）分析，它可以选择性地降低额外的、未环化探针，保留环化的探针。之后，闭锁的探针在两个独立的引物结合域之间的#1位点被切割（初始位置在探针的内部），现在位于被转化探针的一个末端，处于一个适合PCR扩增的位置。然后，PCR扩增4个反应样品中各自的特异性标签，所以当把它们组合到一张芯片上来读出那4个未知碱基时，他们可以清楚地分出各自是哪一种。杂交到芯片上之后（其中包含针对MIP产品上序列片断的额外的探针序列），洗脱各自芯片并使用四色量子点进行染色，然后使用Affymetrix GCS3000四色扫描仪扫描每一个量子点的波长。

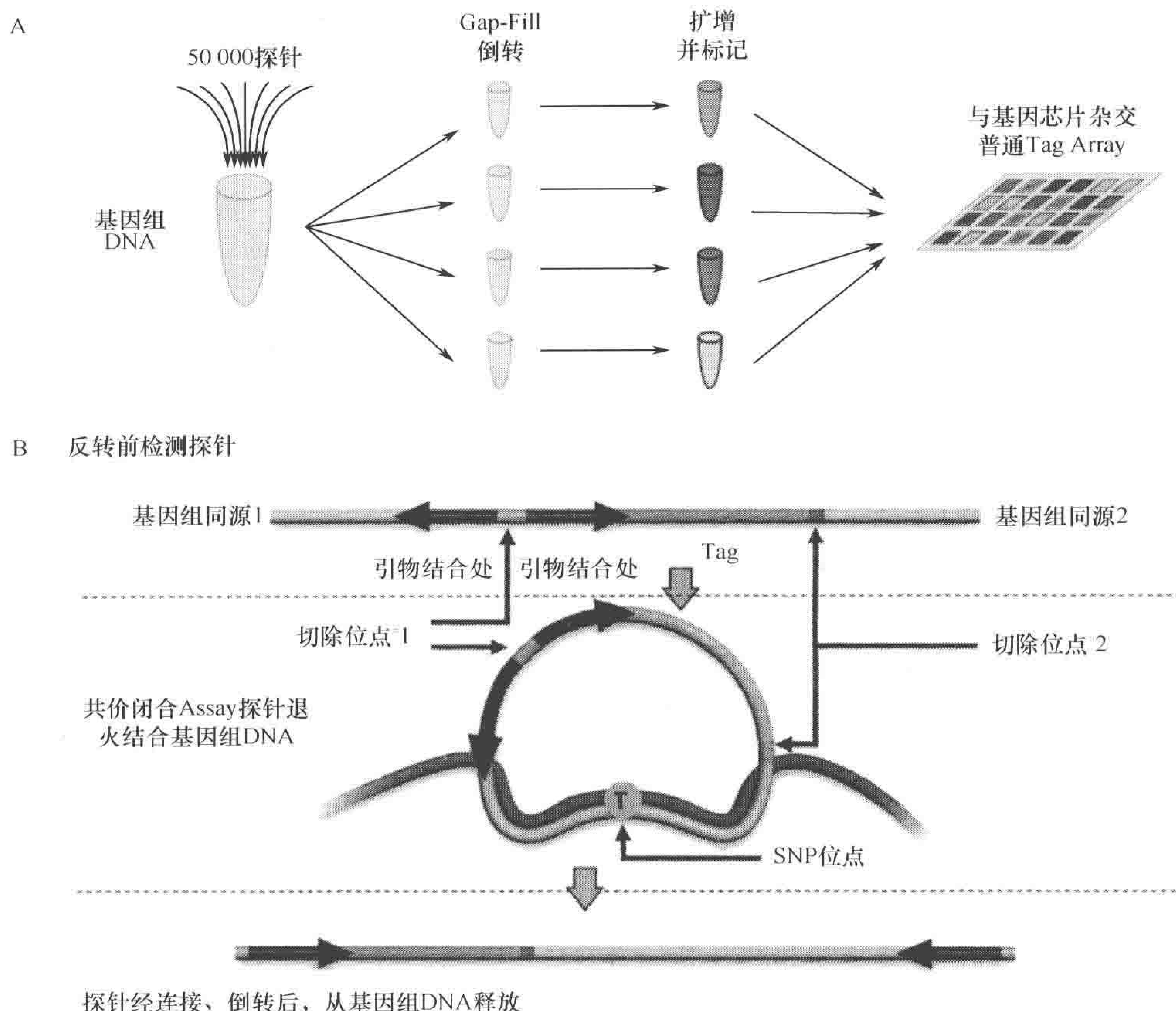



图 12-1　高通量，4 色分子反转探针（MIP）靶向基因分型芯片图解。A. 芯片的制作过程中通过退火将 50 000 探针与样本 DNA 结合，形成芯片 GeneChip 探针点阵杂交。在杂交和洗板后，通过用 Qdot conjugate 荧光染料染色产生基因分型信号，随后用 GCS3000 4 色扫描仪进行扫描。B. 在反转前和反转后进行 MIP。（上图）基因组同源区未反转 MIP 探针的设计。这个基因组同源区包括 SNP 位点，PCR 引物位点，与这个 SNP 相关的标签位点及许多剪接位点。（中图）在'T'SNP 位点退火和连接后，MIP 探针与基因组 DNA 模板相结合。随后，通过在 1＃位点剪接将环状探针反转。（下图）被标记的反转探针用于 PCR 扩增，及形成探针杂交点阵（经作者同意，原图引自 Karlin-Neuman et al. 2007）

在使用了极大化正交合成阵列（Winzeler et al. 1999）解读芯片结果后，MIP 芯片的精确度、敏感性和灵活性进一步得到提高。Affymetrix 芯片的附加标签序列是经过筛选的，具有很接近的熔解温度，显现很低的交叉杂交和最小化的二级结构（Hardenbol et al.，2005）。这些芯片不是专门用于标准阵列，而是可以用作无限种类的产品，每一种都包含期望的生物学标靶序列于一套合适的标签中。为了进一步减少不希望的交叉反应，在 MIP 产品和标签阵列杂交之前，清除了相关的 21-mer 标签之外的基因组 DNA 序列。根据微阵列结构中标签序列的荧光决定位点的基因型。高的信噪比受益于获得一个标记中所有基因型的信息。尽管只有两个潜在的通路可以

给出双等位基因的信号，另外两个背景通道的荧光也可以推测每一个探针在阵列中的特异性［表示为信号-背景（S/B）值］——由于任何原因导致荧光没有标记上的时候，探针将显示低的 S/B 值。最后，每个样品的 4 张芯片图像的大致荧光密度是背景消减，4 种着色光谱重叠校正，并进行标准化。在运用了合适的通过滤波器之后，阵列中的标记确定了基因型。基因型可以被称为无聚类和聚类；前者提供了一种快速、质量控制（QC）样品的结果概要，总结很多有用的规律和图形可视化（下面的“结果”部分）；后者使用了期望最大化（E/M）聚类算法以确认标记聚类位点是纯合型还是杂合型，以及某一样品中的某一标记属于一个特定聚类的可能性（Hardenbol et al. 2005；Moorhead et al. 2006）。这些分析是由 Affymetrix 基因芯片靶向基因分型软件（GTGS）完成的，其中也包含了一个实验室信息管理系统（LIMS）追踪样品芯片处理的整个过程。

## 方案　MIP 靶向 SNP 分型技术和生物芯片

我们使用了一幅简单的插图描述这项技术的效能，其中的数据来源于一张 Affymetrix 25K 芯片，MIP 20 000 错义编码 SNP（cSNP）。该样品同时检测了 20 000 功能性 cSNP 的基因型——该基因编码区域的多态性可以导致氨基酸的改变——覆盖了＞10 000人类基因，SNP 间距＜70kb。选择这些 SNP 的重要证据是它们中的一个大片段有功能性的改变（Ireland et al. 2006）。最近的一项全基因组关联研究使用了这个样品前面一半结果，再次证明了原来怀疑的 4 个 1 型糖尿病易感位点中的 1 个，至少发现了 1 个新位点，即干扰素诱导解螺旋酶（IFIH1）区域，这个结果在其他 6 个不同的民族群体中得到了确认（Smyth et al. 2006）。

我们使用 20 000 cSNP 芯片描述 MIP 基因分型检测技术的步骤如下。图 12-1B 是检测过程中 MIP 探针的变化，图 12-2 是检测流程［相应的步骤见“方法”中的步骤

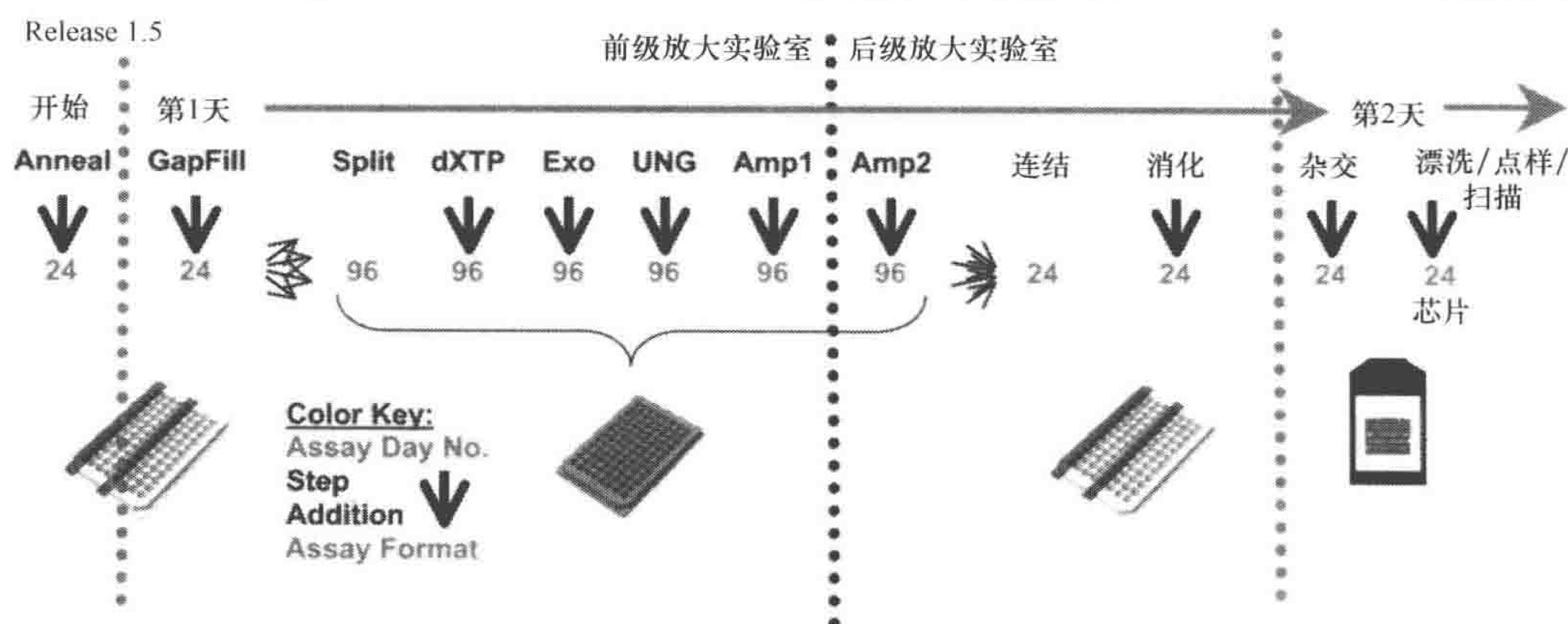



图 12-2　MIP Rel 1.5 芯片的工作流程。24 孔板芯片的流程显示在图中，在第 0 天退火，第 2 天芯片扫描。注意，在过夜退火后，为了检测所有探针（包含任一个碱基 A，C，G 或 T 的样本）形成环形探针的量，在加入 dXTP 以前将这 24 个样本分成 96 孔。将每个样本的 4 个反应，合并后杂交到一个阵列

(5)]。成功使用 MIP 检测技术的关键在于每一步按照说明书小心地混合反应物。如果按照说明书，用良好的分子生物学操作技术处理酶和反应，该检测结果将十分稳定。注意该检测是在两个相对独立的实验室里完成的——前级放大和后级放大——来预防 DNA 扩增产物和先前反应的污染。后级放大实验室的材料、记录和器械不应该带到前级放大实验室，这两类实验室中的实验服应该在各自的实验室中小心地使用。

# 材料

## 试剂

HapMap CEPH DNA Ⅲ。从科里尔研究所获得的这些样品，组成了 CEU 的三体家系，并被用于解释如下描述的 20K cSNP 的数据。这些样品的完整清单可以在 HapMap 网站（www.hapmap.org）上获得。

Taq 聚合酶（Stratagene）

TITANIUM Taq 聚合酶（Clontech/Takara）

## 器材

下列的材料和仪器在 Affymetrix 都可以购买得到，研究者可以在自己的实验室开展 20，000 的 cSNP 基因分型实验。关于其他标准品和定制品的更多信息，或者关于建立在 MIP 技术基础上的器械和基因分型方法的信息，都可以在 Affymetrix 网站上获得（http://www.Affymetrix.com/products/application/targeted_genotyping.affx）。

Affymetrix 20K cSNP 人基因芯片（Panel 1 和 2）。试剂盒包括了足够进行 24 次检查的试剂（包括一个对照）。该产品中包含了 10 000 个确认（双向）的编码功能的错义 cSNP 和 10 000 个 Affymetrix 独立确认。

Affymetrix 25K Tag Array 基因芯片（6-Pack 或者 96-Pack）。芯片上有>25 000 个微孔，使用 Affymetrix 基因芯片 DNA 分析系统结合 MIP 技术每次检测可以确定 20 000个 SNP。

Affymetrix 基因芯片扫描 3000 靶向基因分型系统（GCS 3000 TG 系统），包括

- 计算机工作站和条形码读卡机。前级放大实验室和后级放大实验室样品记录：系统控制。
- 基因芯片杂交加热室 640。加热室可以在 30～60℃ 的杂交温度同时进行 64 个连续的循环。
- 基因芯片清洗系统 FS450（每个体系两个）。每个冲洗系统可以在 30min 内一次性自动化清洗和染色 4 张芯片。
- 基因芯片扫描仪 3000 7G 4C。四色共聚焦激光扫描仪包含了一个支持 48 张芯片的温度敏感自动装卸机。芯片在基因芯片控制系统软件（GCOS）v1.4 控制下自动扫描和记录。
- 基因芯片 TG 分析软件 v1.5。分析扫描文档并产生基因分型结果。

# 方法

这种策略和分型步骤的观点总结如下（图 12-2）。为了用更多的解释来优化不同的复杂水平完善这种技术的细节，就产生了基因芯片靶向分型系统的用户手册。

## 前级放大实验室工序

（1）样品 MIP 探针退火（退火）。用 4μg DNA 样品结合芯片，用酶 A 处理。DNA 变性（95℃，5min），58℃下孵化 16～24h（但是要<30h）。

（2）在样品 DNA 中加入退火探针（Gapfill，Split，dXTP）。

i. 在模板中加入连接酶和 4 种连接反应物（A、C、G、T）。58℃下孵化 10min。

ii. 在每个反应中加入合适的单一 dXTP，58℃下孵化 10min。

（3）消除未环化探针（Exo）。接入核酸外切酶混合物，在 37℃下孵化 15min。加热灭活加入的酶（95℃，5min）。

（4）裂解，插入 MIP 探针（裂解）。加入 UNG 裂解混合物，在 37℃下孵化 10min。

（5）使用通用 PCR 引物扩增插入的 MIP 探针（Amp1）。加入 Amp1 混合物，用 Meg 20K 1 st PCR 热循环体系孵化 20 个循环（运行 1h）。

在这一步可以运行质量控制，在进入下一步之前来评估产物的质量。

## 后级放大实验室工序

（6）碱基-（或者系统-）特殊标签的扩增 MIP 产品（Amp2）。

i. 加入少量（4μL）的各种 Amp1 反应物（A、C、G、T）到包含合适的等位基因特异性引物的 Amp2 混合物中。

ii. 用 Meg Hypcr 10～20K 和 PCR 体系扩增 10 个循环（运行 30min）。

（7）在相关片断中消除基因组 DNA（消化）。每个样品和 6μL 消化混合物结合 4 种反应物（A、C、G、T）中的一半，在 37℃下消化 1.5h。加热灭活酶（95℃，5min）。在这一步可以运行质量控制，在芯片分析之前来评估产物的质量。注意在变性前必须移除用于质量控制的样品。

（8）杂交 MIP 产品（用合适的体系把等位基因转化为片段）到芯片（hyb）。混合杂交混合物和消化过的 MIP 产品，变性，在 39℃下杂交芯片 12～16h（每个样品一张芯片）。

（9）清洗，染色，扫描芯片（清洗/染色/扫描，Ⅰ-Ⅲ自动运行）。

i. 可选择的：手工移除芯片中的杂交混合物（如果有必要，就保存）。

ii. 顺序在低-、高和标准情况下，清洗系统每次可以清洗 4 张芯片。

iii. 用 Qdot 染色混合物染色芯片，缓冲液重新充满。48 张芯片的清洗和染色需要 3h。

iv. 把芯片装载到自动装卸机上，进行扫描。扫描每个样品的 4 个文档（A、C、

G、T）和芯片，4 个系统需要 6min。每个样品产生 4 个 cel 文件（A、C、G、T），每个 cel 文件都包含 tag 序列的初始荧光信号。

（10）分析数据。输入每个样品的 Cel 文档和记录的样品信息到 GTGS 软件，进行系统和芯片的标准化，消减背景、光谱的色度干扰。进行前级和后级聚类分析，使用计算机分析，输出聚类基础上的基因分型。从数据库获得的大量数据自动化分析（下面的结果部分）。

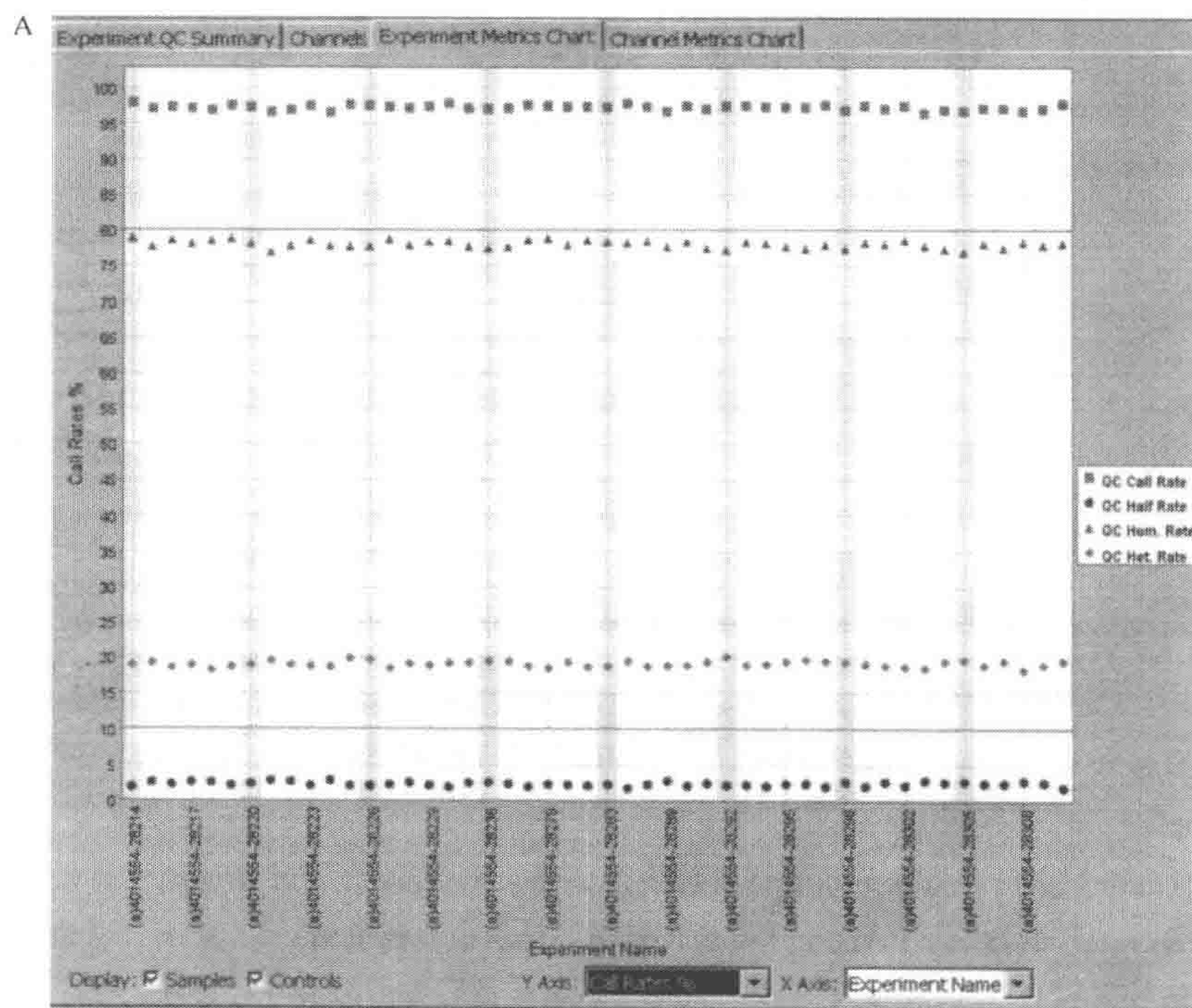
## 具有代表性的数据分析和 20K cSNP 芯片操作的可视化

在这个例子中，48 个 CEPH HapMap 样本，包括 5 个三体家系和 3 个重复样本，在 MIP 20K 标准芯片状况下运行 20K cSNP 芯片。扫描的初步数据输出到 GTGS 分析软件，使用标准的参数来消除任何可疑的探针。数据过程的细节参见 Moorhead 等（2005）。

### 前聚类数据分析和 QC（质量控制）结论要点

QC 简要分析提供了一种快速的方法来评估样本的性能，提供了表格和绘图的形式。经过了合适的标准化、背景消减和光谱交叉校正，标记经过各种标准（如信号最小化、S/B、S/N、等位基因系统信号比例），按照等位基因 1 或 2 评为纯合子或者杂合子。信号的总数目，表述为芯片中的百分数，报告为样品的“QC 信号比”，在散点图上标出每个样品的结果。图 12-3A 显示了每个 CEPH 样品的信号比，代表“Experiment

A



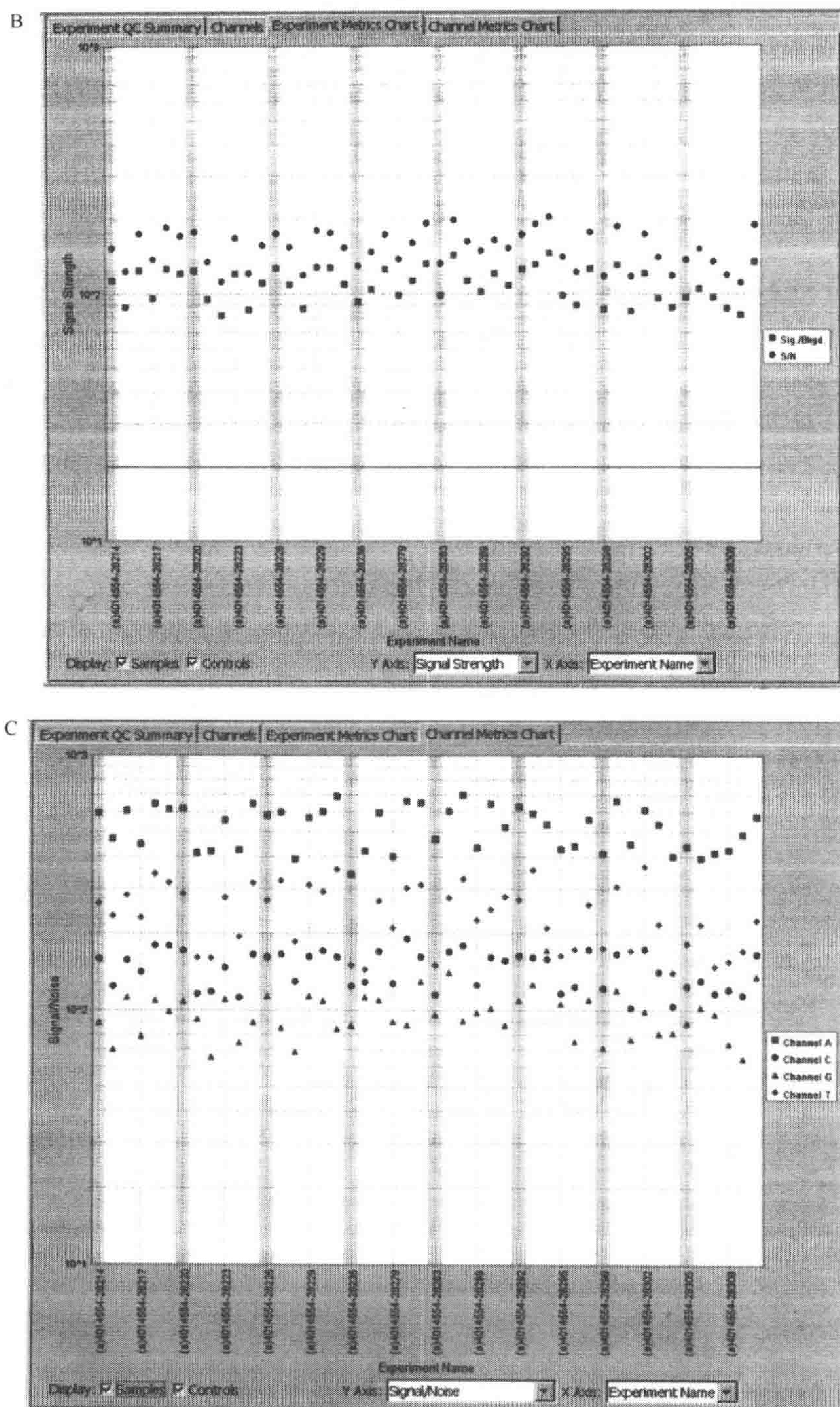



图 12-3　（继续）通过 GTGS 分析软件进行质量控制。A. 这个“实验计量图”显示了质量控制应答率（call rate）的数值。B. 这个“实验计量图”显示信号强度值。C. 这个“通道计量”图显示信号噪声值（S/N）比率。没有其他“通道计量”显示的信号、背景、噪声等值

Metrics Chart”的观点。QC 信号比在 80％以下的自动淘汰（显示在这个位置的线）。这 48 个样品的 QC 信号比，纯合子和杂合子的常频落在约 97％和 98％（分别为约 78％和 20％）。剩下的一些不清楚的信号被记为半信号，意味着一个等位基因确定而另一个不确定。聚类（见下述），很多这些半信号可以清楚地解决。这种观点里的其他 Metrics 是“信号说服力”（图 12-3B）。这种观点揭示了芯片的信噪比（S/N，蓝色符号），以每个通道平均值为基础的 4 通道平均值修正了所有标记的荧光信号和通道的噪声；信号-背景比值（S/B，红色符号），两个通道等位基因信号的平均值与两个通道背景值（或者没有等位基因）的比值。为确定基因分型的最小化 S/B，设置蓝线的 S/B＝20；这个限制是远远大于这些样品的，它们大多接近或超过 S/B＝100。

同样的，四通道中的大量情况被描记在“Channel Metrics Chart”（图 12-3C），还可以用来追踪和诊断任何问题。包括信号、噪音、S/N、芯片背景、舍弃异常值。图 12-3C 显示了 S/N 的情况，看得出合理的一致性和特定通道的特性。当一个样本在运行后没有成功，与其他样本的 Metrics 对比，可以发现是否问题源于其中的一个通道、一些不常见的芯片高背景或可能不足/低质量的样本 DNA 输入。为了使诊断更加方便，在每个芯片和通道使用额外的绘图方法（未展示），可以允许评估生信号和熟信号的区别，在不同的转变中控制实验的样本。

## 聚类分析和结论要点

QC 简要分析之后，用户接着要根据默认值或可变参数值进行聚类分析。期望值/极大值化（E/M）聚类算法使用产生基因型，输出不同的可选择的格式。GTGS 分析软件产生一个如图 12-4A 所示的对 48 个 CEPH 样本 20K cSNP 芯片分析的基因型聚类简要表格。注意所有的 48 个样本（或“实验”）是成功分型的（比较“＃实验”和“＃分型实验”），芯片中的 20127 个标记（＃“检验”），样本的 80％中，99.18％是成功的（“成功的检验％”）。（如果一个标记或芯片的信号在 80％的样本中是成功的，这个样本体系是失败的）回顾 QC 简要中信号比低于聚类 1％～2％的估计。检验的精确度 99.92％（聚类重复符合％），基于该 48 个样本三次重复 HapMap（“＃单一样本基因分型”＝45 对“＃基因分型实验”＝48），准确度为 99.91％（“三体家系聚类一致度％”）。基于那 48 个样本中的 5 个父-母-子三体家系（“＃三体家系”）的基因型传递。“完整性％”＝99.51％显示 48 个样本中的 19，962 个通过的检测，仅有 0.49％（1-完整性）的数据丢失，例如，几乎所有样本中的几乎每一个标记都有信号。其他的数据和标记无信号的原因也被记录。

通过 QC 简要的数据，基因型分型的数据聚类分析可以在芯片水平（“Experiment”键，未显示）看到每个单一样本中所有标记的情况，或者标记水平（“Assay”键）可以看到每一个个体的标记在所有样本中的聚类。图 12-4B 显示了聚类分析和加亮的代表性的 A/C 标记（rs9353689）在“Assays”键中的表现。在这个一维图中，A/A 纯合子位于“－1.00”或其附近；C/C 纯合子位于“1.00”或其附近；A/C 杂合子位于“0.00”或其附近。Y 轴显示的是两个等位基因通道信号的总数。实线延伸交叉于每个聚类代表平均聚类位置±3$\sigma$。一个样品中的一个标记是否可以成为特殊聚类的成分取决

于它变色于那个分类还是另外的，可以与其附近的聚类区分彼此。如该分析窗口所示，三体家系成员的基因型包括了孩子的样本，NA10846，标记为虚竖线。我们发现三体家体中父亲与母亲是相对的纯合子（母亲＝A/A，父亲＝C/C），孩子的基因型期望值应该是杂合子（中间聚类）。我们可以在标记表格中选择为任意需要测定的 SNP 画图，

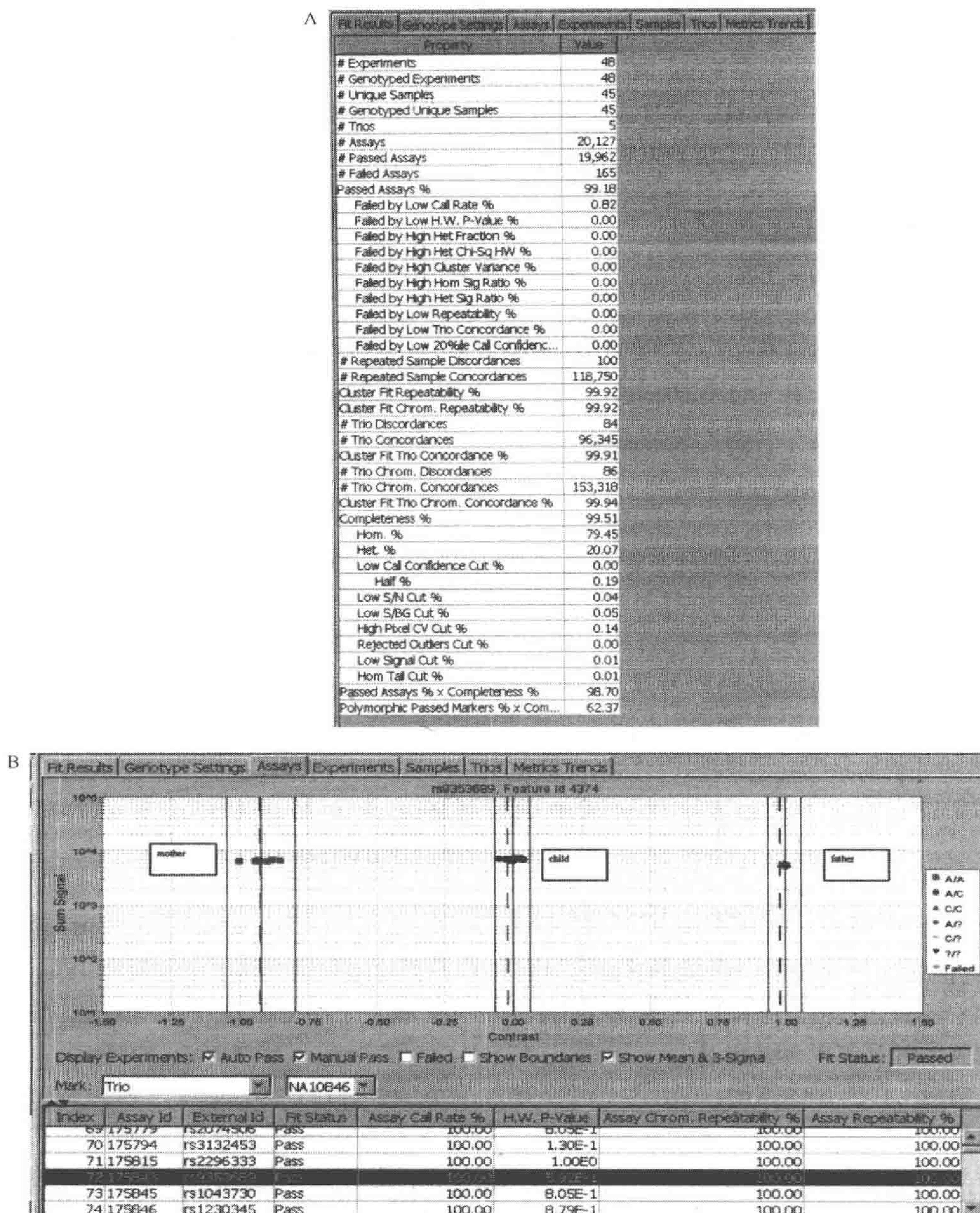



图 12-4　通过 GTGS 分析软件做的聚类分析。A. 这个图显示聚类适合度总结表。B. 这个图显示的是阵列标记聚类图。详细解释见正文

表中每一个标记的很多其他参数也是可见的（如分析信息率、可重复性、Hardy-Weinberg P-值、样本中次要等位基因频率、染色体和染色体位置、基因位点）。其他选择键显示使用样本的信息和他们在样本系统中的频率（samples）、三体的情况数据（trios）、实验或者样本表现数据（metrics trends）和聚类分析参数值（genotype settings）。

## 摘要和总结

基于 MIP 的靶向基因型分析是一种用途广泛、高通量、复杂的检测，可以在样本 DNA 中同时分析 1500 到＞50 000SNP 标记，用于各种用途。结合了特异酶和内在特异探针的靶向识别使得 MIP 技术具备高精度、可重复性和数据的完整性。另外，这些特征允许最小的设计局限性和检测探针对大多数靶向序列的高转化率。MIP 基因分型技术的多功能性，由于成功地对更多有挑战性的目标序列进行分型被进一步证实，如包含了插入/缺失、三等位基因和继发 SNP（Dumaual et al. 2007）。在芯片结果读出时使用合成杂交标记，不仅提高了芯片的正确度，而且允许芯片的成分可变——这可以为同一个检测制造标准或定制芯片，通过增加未使用的阵列标签还可以增加芯片的种类。最后，一个丰富而直观的软件分析系统（GTGS 分析软件）处理检测数据产生聚类导向基因型，以及很多参数和图表，可以通过一项研究来增加个体标记和样本的情况。

## 致谢

我们感谢 ParAllele、Affymetrix 和很多在 MIP 技术发展中尽力奉献的人。这些贡献来自各个方面，包括销售团队为明确检测的需要和能力所作的贡献；信息团队为探针设计、算法发展和最终软件包所作的贡献；制造探针和反应物的贡献；化验进展团队；检测服务团队对数据产生和分析所作的贡献。

## 参考文献

Begovich A.B., Carlton V.E., Honigberg L.A., Schrodi S.J., Chokkalingam A.P., Alexander H.C., Ardlie K.G., Huang Q., Smith A.M., Spoerke J.M., et al. 2004. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **75**: 330–337.

de Bakker P.I., Yelensky R., Pe'er I., Gabriel S.B., Daly M.J., and Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* **37**: 1217–1223.

Dumaual C., Miao X., Daly T.M., Bruckner C., Njau R., Fu D.J., Close-Kirkwood S., Bauer N., Watanabe N., Hardenbol P., and Hockett R.D. 2007. Comprehensive assessment of metabolic enzyme and transporter genes using the Affymetrix® Targeted Genotyping System. *Pharmacogenomics* **8**:293–305.

Hardenbol P., Yu F., Belmont J., MacKenzie J., Bruckner C., Brundage T., Boudreau A., Chow S., Eberle J., Erbilgin A., et al. 2005. Highly multiplexed molecular inversion probe genotyping: Over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* **15**: 269–275.

International HapMap Consortium 2005. A haplotype map of the human genome. *Nature* **437**: 1299–1320.

Ireland J., Carlton V.E., Falkowski M., Moorhead M., Tran K., Useche F., Hardenbol P., Erbilgin A., Fitzgerald R., Willis T.D., and Faham M. 2006. Large-scale characterization of public database SNPs causing non-synonymous changes in three ethnic groups. *Hum. Genet.* **119**: 75–83.

Karlin-Neumann G., Sedova M., Falkowski M., Wang Z., Lin S., and Jain M. 2007. Application of quantum dots to multicolor micorarray experiments: Four-color genotyping. *Methods Mol. Biol.* **374**: 239–251.

Moorhead M., Hardenbol P., Siddiqui F., Falkowski M., Bruckner C., Ireland J., Jones H.B., Jain M., Willis T.D., and Faham M. 2006. Optimal genotype determination in highly multiplexed SNP data. *Eur. J. Hum. Genet.* **14:** 207–215.

Nilsson M., Malmgren H., Samiotaki M., Kwiatkowski M., Chowdhary B.P., and Landegren U. 1994. Padlock probes: Circularizing oligonucleotides for localized DNA detection. *Science* **265:** 2085–2088.

Risch N. and Merikangas K.1996.The future of genetic studies of complex human diseases. *Science* **273:** 1516–1517.

Shiffman D., Ellis S.G., Rowland C.M., Malloy M.J., Luke M.M., Iakoubova O.A., Pullinger C.R., Cassano J., Aouizerat B.E., Fenwick R.G., et al. 2005. Identification of four gene variants associated with myocardial infarction. *Am. J. Hum. Genet.* **77:** 596–605.

Smyth D.J., Cooper J.D., Bailey R., Field S., Burren O., Smink L.J., Guja C., Ionescu-Tirgoviste C., Widmer B., Dunger D.B., et al. 2006. A genome-wide association study of nonsynonomous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nat. Genet.* **38:** 617–619.

Van Eerdewegh P., Little R.D., Dupuis J., Del Mastro R.G., Falls K., Simon J., Torrey D., Pandit S., McKenny J., Braunschweiger K., et al. 2002. Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* **418:** 426–430.

Winzeler E.A., Shoemaker D.D., Astromoff A., Liang H., Anderson K., Andre B., Bangham R., Benito R., Boeke J.D., Bussey H., et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:** 901–906.

Zheng S.L., Liu W., Wiklund F., Dimitrov L., Balter K., Sun J., Adami H.O., Johansson J.E., Chang B., Loza M., et al. 2006. A comprehensive association study for genes in inflammation pathway provides support for their roles in prostate cancer risk in the CAPS study. *Prostate* **66:** 1556–1564.

# 互联网资源

**http://www.affymetrix.com/products/application/targeted_genotyping.affx** Affymetrix targeted genotyping page

# 13 全基因组基因分型

Stacey B. Gabriel[1] and Michael P. Weiner[2]

[1]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142;* [2]*RainDance Technologies, Guilford, Connecticut 06437*



## 简介

在满足进行全基因组关联扫描的质量和成本效率要求的前提下，有几种技术能够对大样本检测，得出成千上万个多态性位点的数据。因为全基因组分析需要非常大的财力支持，所以已经发展为商品化的全面解决方案（turn-key）。只有到全基因组 DNA 测序在价格方面具有竞争优势时，才可能广泛应用寡核苷酸阵列进行全基因组研究，就像现在这样。基于 DNA 的阵列制备方法有很多种，包括原位化学合成法（Affymetrix，Nimblegen）、压电喷头（或喷墨喷头）喷样法或者无序自组装微球阵列法（Lynx，Illumina）（Walt 2002）。在下面的几节，我们将介绍两种基于 DNA 阵列的商业化全基因组分析方法。

## 应用 Affymetrix 基因芯片进行高密度基因组变异分析

### 基因芯片的发展史

2000 年起，Affymetrix 就开始投资研发用于单核苷酸多态性（single-nucleotide polymorphism，SNP）基因分型的基因芯片。第一张芯片密度为 10K，后来陆续发展成 100K（2003）、500K（又称为 SNP 5.0），最近又推出了 1M 的 SNP 检测芯片（又称为 SNP6.0）。

产品发展过程中基本的基因分型方法并没有改变，但是随着 Affymetrix 降低由脱氧寡核苷酸探针决定每个样品点尺寸的能力不断提高，阵列的密度得以不断增大。这是基于约化表示基因分型的一种方法，通过酶消化和片断选择降低基因组的复杂性（Altshuler et al. 2000）。通用 PCR 扩增，标记 PCR 产物标记后杂交到 Affymetrix 阵列上（Kennedy et al. 2003；Matsuzaki et al. 2004），这些技术的使用降低了基因组的复杂性。利用限制性酶消化，确定限制性片段内相应 SNP 位点的基因型。Affymetrix 开发了一种算法，能够计算出检测所需的限制性酶和可能的 SNP 的最佳分布。

Affymetrix 遵循常规方法发展 SNP 阵列容量（图 13-1）。首先进行计算机模拟筛选：根据 SNP 的密度和间距选择限制性酶切片段组合，该组合包含最佳 SNP 含量。SNP 5.0 和 SNP6.0 芯片都选择了 NspⅠ、StyⅠ这两种酶，使复杂的基因组全长缩减至 500Mb，包含 250 万个独立的 SNP（来自 dbSNP）。SNP6.0 芯片在对 Haplotype Map（HapMap）样本的筛查实验中，发展到一个筛查芯片筛查 200 万个 SNP。这 200 万个 SNP 不同于已有的 SNP5.0 芯片上的 50 万个 SNP。在这一筛查阶段，如出现非特异扩增，显示其他杂交问题（如与其他区域高度相似，或者探针特异性杂交出现异常），或者对特殊类别基因型识别能力差等情况（检出结果与 HapMap 做一致性比较，作为筛选程序），均能够被识别。实验性筛查完成后，经筛查，screening 芯片能成功检测出 800 000 个 SNP。因为 SNP6.0 芯片最后的 SNP 芯片包括 SNP5.0 芯片的 500 000 个 SNP，Affymetrix 又从通过筛查得出的其他 SNP 中，选择了在 HapMap 上最有代表性的一些 SNP，与 SNP5.0 芯片上的 SP，共同组成最佳 tag-SNP（关于 tag-SNP 的讨论部分，见第 20 章）。总计，SNP6.0 芯片包含 907 000 SNP。

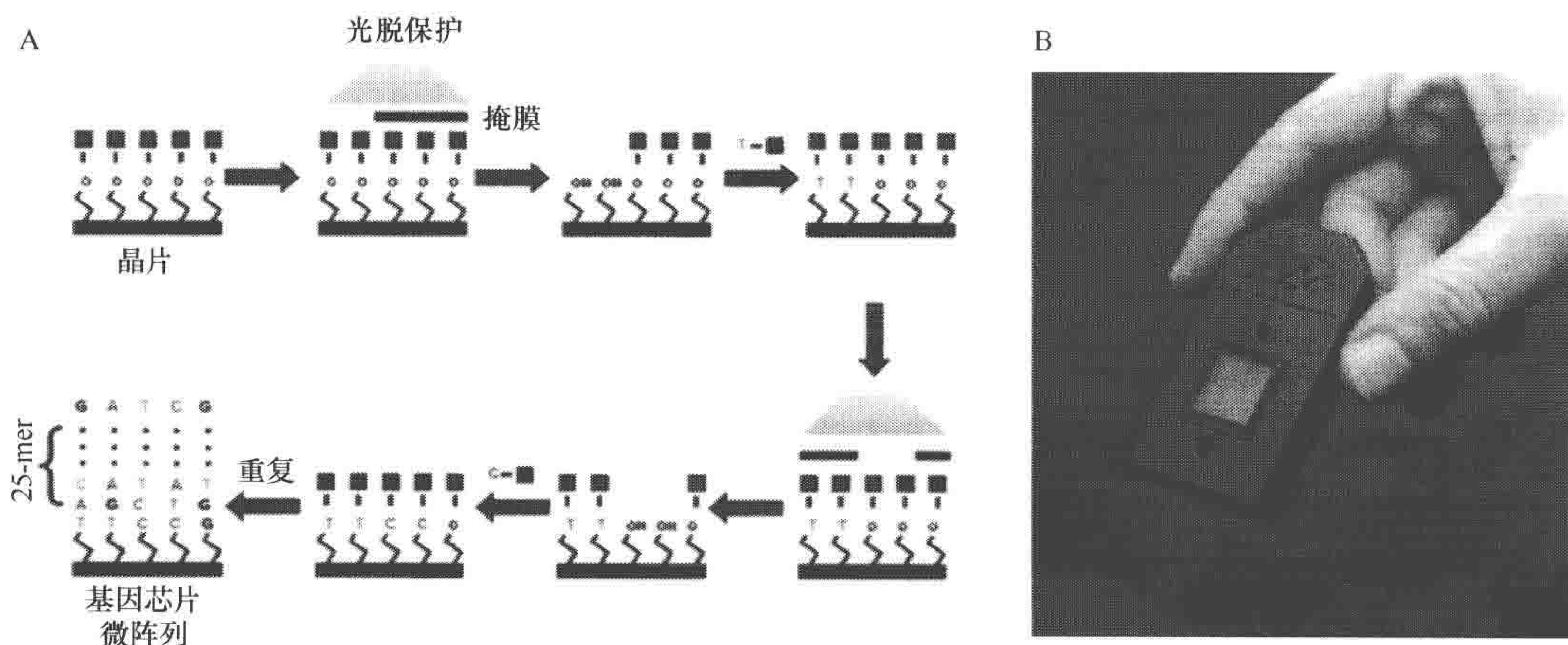



图 13-1　基因芯片的制作方法

Affymetrix 的分析方法和产品开发策略决定其产品将不偏向任何特殊人群或 SNP 芯片，覆盖范围包含一定程度的冗余。这种方法的优势在于不同的患者群体间遗传力应该相似，HapMap 多标记单倍型方法能增大遗传覆盖率，并且因为该产品具备最高的可检测 SNP 密度（平均每 2.5kb 1 个 SNP），使其拥有最佳的总体解决方案用于检测拷贝数变异。主要的缺点是在 SNP 选择上相对不够灵活，因为这种方法是通过限制性内切

酶降低基因组的复杂度。然而，对于任何产品标准，在大规模生产和可预测的市场需求面前，在最终的灵活性和低成本之间都存在一种权衡。

在写这篇文章的时候，SNP 5.0 在该领域应用时间最长。有结果显示，有少数 SNP 在该芯片上得不到重复；大约 470 000 个 SNP 在多项研究中结果持续可信。相对于 HapMap 的检出符合率和精确度作为产品技术性能的评价指标。根据 Affymetrix 的数据，通过与 HapMap 作一致性比较计算出的准确度达 99.5%，检出基因型重复性为 99.9%；这些数据与许多使用该产品进行的全基因组关联研究发表的数据一致（Diabetes Genetics Initiative 2007）。

## 遗传覆盖率

全基因组基因分型产品的遗传覆盖率通常是用 HapMap 上 SNP 所占比例来描述的，这样能使基因分型芯片上所选 SNPs 更有代表性（Barrett and Cardon 2006；Pe'er et al. 2006）。通过报告强相关（$r^2>0.8$）HapMap SNP 所占比例或者评估 HapMap SNP 与芯片上 SNP 的平均相关性来衡量。后者统计结果与相关研究中的效能相符，因为前者使用了一定的相关系数阈值，而相关系数小于 0.8 的 SNP 在大样本相关研究中还能提供信息。Affymetrix 5.0 和 6.0 最大 $r^2$ 平均值介于 0.68～0.90，依据产品和 HapMap 受检人群不同而不同（表 13-1）（The International HapMap Consortium 2005）。

**表 13-1 各种基因分型平台的比较**

| Platform | HapMap Panel | | | | | |
|---|---|---|---|---|---|---|
| | YRI | | CEU | | CHB+JPT | |
| | $\%r^2 \geq 0.8$ | mean max $r^2$ | $\%r^2 \geq 0.8$ | mean max $r^2$ | $\%r^2 \geq 0.8$ | mean max $r^2$ |
| Affymetrix SNP5.0 | 46% | 0.66 | 68% | 0.81 | 67% | 0.80 |
| Affymetrix SNP6.0 | 66% | 0.80 | 83% | 0.90 | 81% | 0.89 |
| Illumina HumanHap300 | 33% | 0.56 | 77% | 0.86 | 63% | 0.78 |
| Illumina HumanHap550 | 55% | 0.73 | 88% | 0.92 | 83% | 0.89 |
| Illumina HumanHap650Y | 66% | 0.80 | 89% | 0.93 | 84% | 0.90 |

注：YRI. 约鲁巴人；CEU. 高加索人；CHB. 中国北京汉族人；JPT. 东京日本人。

# 应用 ILLUMINA BEADCHIPS 进行高密度基因组变异分析

## 微阵列技术发展史

微珠组装是将结合寡核苷酸引物的微珠装入光纤面板，制成高密度阵列。纤维孔最初是用一种能特异腐蚀纤维孔的酸局部蚀刻而成（而不是光学熔覆，图 13-2）。孔的深度由酸的浓度、反应时间和温度控制。包层和底部的孔壁是由凹进去的光纤末端形成，进而形成微孔。当蚀刻微孔被浸在乳胶或二氧化硅微珠溶液中，微珠与孔的大小正好相符，微珠借毛细作用组装进到微孔里。为了更好地质控，各种微珠是成批制备出来的。

将不同的寡核苷酸引物组合固定在微珠上，作为光学“条码”，使其结合的微珠得以解码。不同的微珠组成一个文库，然后使微珠组装进入一个纤维阵列的刻蚀孔中。因

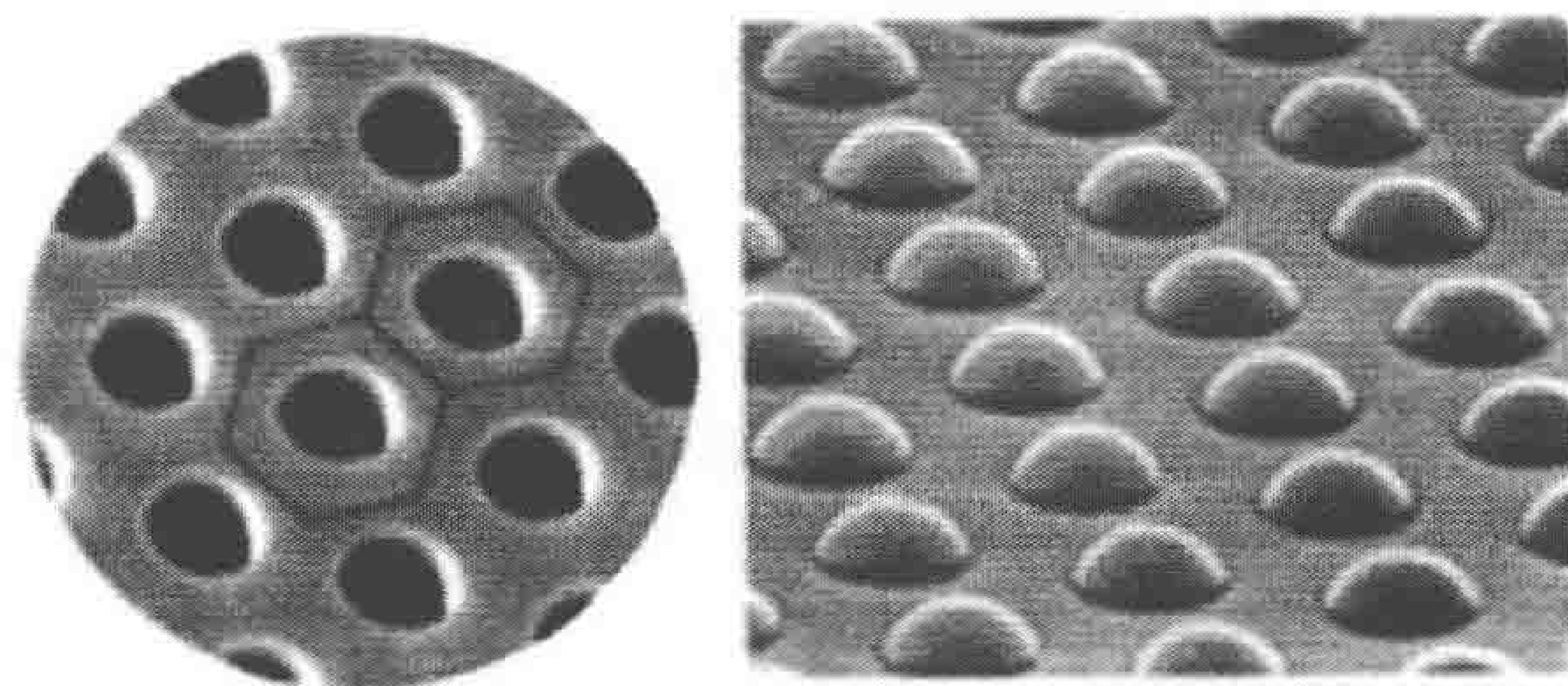
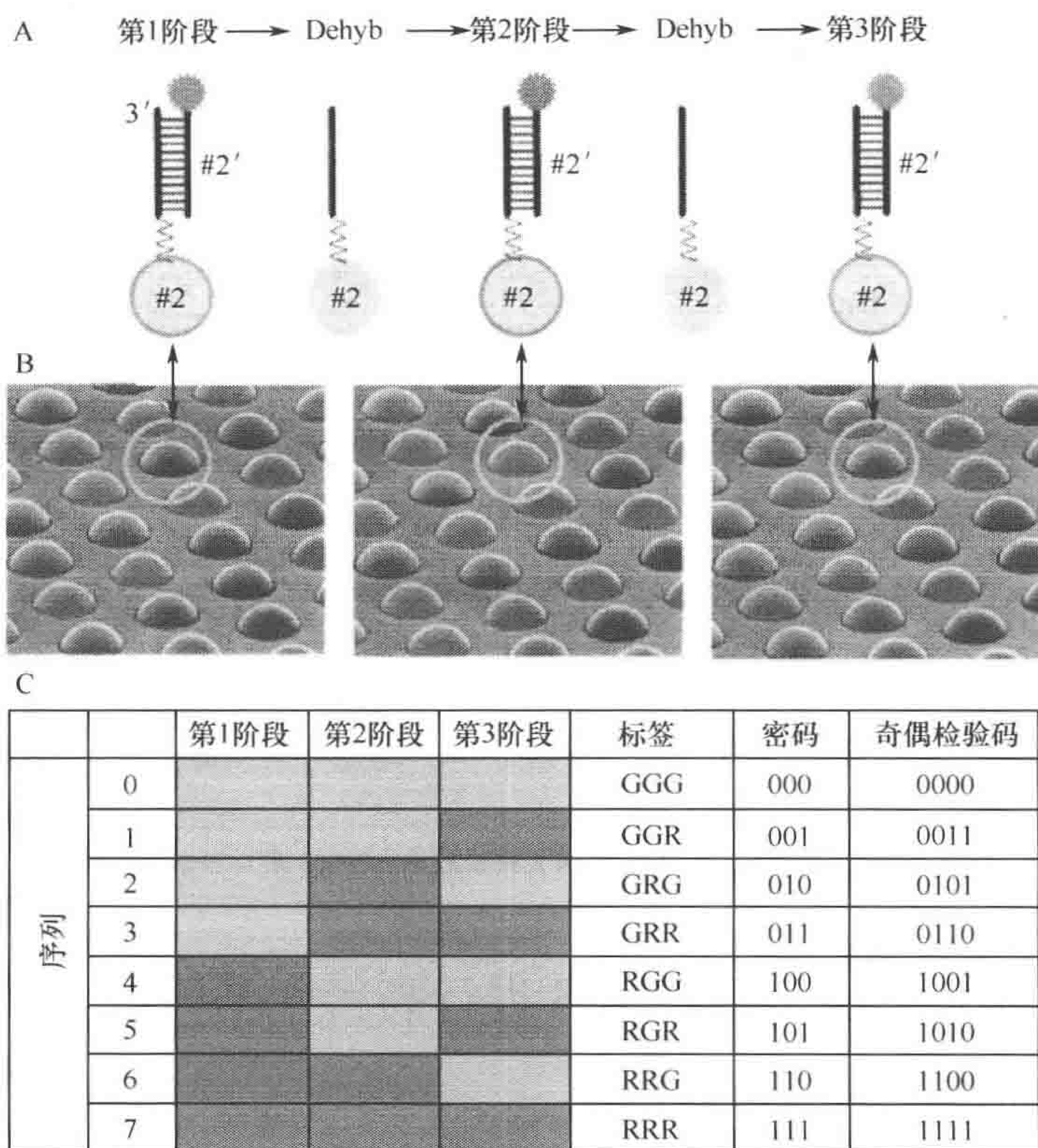


图 13-2　光纤微珠阵列（左）光纤面板上的微孔呈六角密排结构，微孔是由酸选择性腐蚀芯玻璃而成，光纤包层玻璃不受腐蚀，相对完好。（右）蚀刻面板负荷着编码微珠

为微珠无序自组装进入各个孔中，所以条码起到记录每种类型微珠位置的作用。DNA 阵列包含数千个连接在光学纤维末端的微珠。

Illumina 为每个微珠设计独一无二的寡核苷酸序列，并结合到微珠上，用于解码阵列上随机组装的微珠（图 13-3）。每个微珠类型平均被使用 30 次，每次实验数据取平均




| | | 第1阶段 | 第2阶段 | 第3阶段 | 标签 | 密码 | 奇偶检验码 |
|---|---|---|---|---|---|---|---|
| 序列 | 0 | | | | GGG | 000 | 0000 |
| | 1 | | | | GGR | 001 | 0011 |
| | 2 | | | | GRG | 010 | 0101 |
| | 3 | | | | GRR | 011 | 0110 |
| | 4 | | | | RGG | 100 | 1001 |
| | 5 | | | | RGR | 101 | 1010 |
| | 6 | | | | RRG | 110 | 1100 |
| | 7 | | | | RRR | 111 | 1111 |

图 13-3　微珠类型解码过程。A. 2 型微珠的单个微珠顺序杂交过程图解。第 1 阶段，互补的解码序列与微珠上固定的寡核苷酸探针杂交。解码序列标记有荧光基团（第 1 阶段为绿色，第 2 阶段为红色，第 3 阶段为绿色）。将整个阵列成像读取荧光信号。再将阵列去杂交，此过程重复两次以上。B. 一张微珠阵列的扫描电镜显微照片，经人工着色，分别代表三个顺序杂交阶段。综合比较，这些图像反映出各个微珠的组合代码。注意黄色圆圈的微珠颜色特征为 GRG，代码为 010。C. 将不同的颜色或状态分配给各反应阶段的每个解码序列，整个阶段下来则产生唯一的颜色组合。颜色特征或代码能确定微珠类型。如奇偶纠错码这一栏所示，增加额外解码阶段（数据为给出）能提供纠错奇偶校验位。经过三个解码阶段，所有微珠根据颜色被唯一识别（经允许，转载自 Gunderson et al. 2004.）。（见图版）

值。高度冗余的数据集产生高质量的数据。现在，Illumina 生产出多种规格的高密度 SNP 基因分型芯片（“BeadChip;”见综述，Shen et al. 2005），SNP 容量从 109K (Sentrix Human-1，SNP 多以外显子为中心设计）到 650 K 不等。

Infinium 检测法能对人类基因组变异进行大规模筛查，研究者一次就能经济有效的分析整个基因组。实验操作仅需单管处理样品，无需 PCR 扩增或者连接反应等，明显降低了劳动力要求和样品处理过程中发生潜在错误的可能性。酶识别过程（见下述）确保高检出率和精确性。

## HapMap 芯片

tag-SNP 是能够代表许多其他 SNP 的多态位点。tag-SNP 的使用显著提高了相关研究的能力。因为 tag-SNP 已经被应用于人类单体型图的绘制，与其他使用大量随机选取 SNP 的策略相比，使用较少的 SNP 和统计检验即能得到更大的统计能力和基因组覆盖率。317K 的 BeadChip 包含 300 000 个以上 tag-SNP 位点，这些 tag-SNP 取自 HapMap 计划，各 SNP 位点平均间距 9kb。Illumina HumanHap-1 基因分型微珠芯片是 Illumina 能对全基因组进行基因分型的容量最大的产品。与 Affymetrix 降低基因组复杂度相反，Illumina 在等位基因鉴别之前先将整个基因组扩大。于是，Illumina 能够添加在计算机模拟设计过程中漏掉的任何一个 SNP。

InfiniumⅡ检测法是一种基于全基因组扩增后 BeadChip 杂交的方法。用于 SNP 定位的脱氧寡核苷酸探针与光纤面板上的微珠杂交。当扩增的基因组 DNA 与微珠阵列特异性杂交后，阵列上的等位基因特异性引物延伸，双色荧光标记，根据微珠上检测到的荧光颜色能够确定等位基因类型。

Illumina 产品的含量在很大程度上取决于 HapMap 上的标签信息，通过连锁不平衡选择算法，将 SNP 分成若干子集（第 19 章）。从各子集中选择 tag-SNP，经预测在 Illumina 平台上最可能成功检测的 SNPs 优先入选。作为标签的标准，入选子集的 tag-SNP 必须满足在不同距离的基因（根据不同产品确定）或进化保守区内 $r^2>0.8$，并且在该区域外阈值要相对低一点。Illumina 还专门增加了选自 dbSNP 的 8000 个错义和主要组织相容性复合体（MHC）区域的 1500 个 SNP。Illumina 的三种规格的产品，分别分析 317 000 个（见前文）、550 000 个和 650 000 个 SNP。前两种被选择用于 CEU HapMap panel（高加索人），密度最高的芯片已经增加 SNP 含量以覆盖 YRI HapMap panel（Yoruba）。关于覆盖率的统计见表 13-1。

## 样品扩增与杂交

全基因组扩增需要投入 250～750 ng 基因组 DNA（g DNA），因为 Infinium 法单一 BeadChip 即需要大量的 DNA（1000 倍扩增）。扩增完成后，将产物片段化，用异丙醇沉淀（加沉淀试剂），在含有甲酰胺的杂交缓冲液中重悬。DNA 样本在 95℃变性 20min 后装入 Tecan 流动小室，置于湿化瓶中使 SNP 位点杂交上 50-mer 的捕获探针（Steemers and Gunderson 2005；Steemers et al. 2006）。杂交后，将 BeadChip/Tecan 流动小室组件放在控制温度的流动架上，通过向流动小室添加各种试剂即可完成随后的洗脱、

延伸、染色等步骤。

对于等位基因特异性引物延伸（allele-specific primer extension，ASPE；Infinium Ⅰ）法（图 13-4），首先洗去未杂交和非特异杂交的 DNA，然后在加入延伸反应体系之前封闭 BeadChips。在延伸过程中，与 BeadChip 上杂交的 DNA 正确配对的探针得以延伸，并掺入生物素标记的核苷酸。延伸反应结束后，用甲酰胺洗脱，除去杂交的 DNA，以降低无关信号。芯片再经多层染色处理，使信号放大，检测掺入的标记。最后，将 BeadChips 漂洗、干燥、成像。

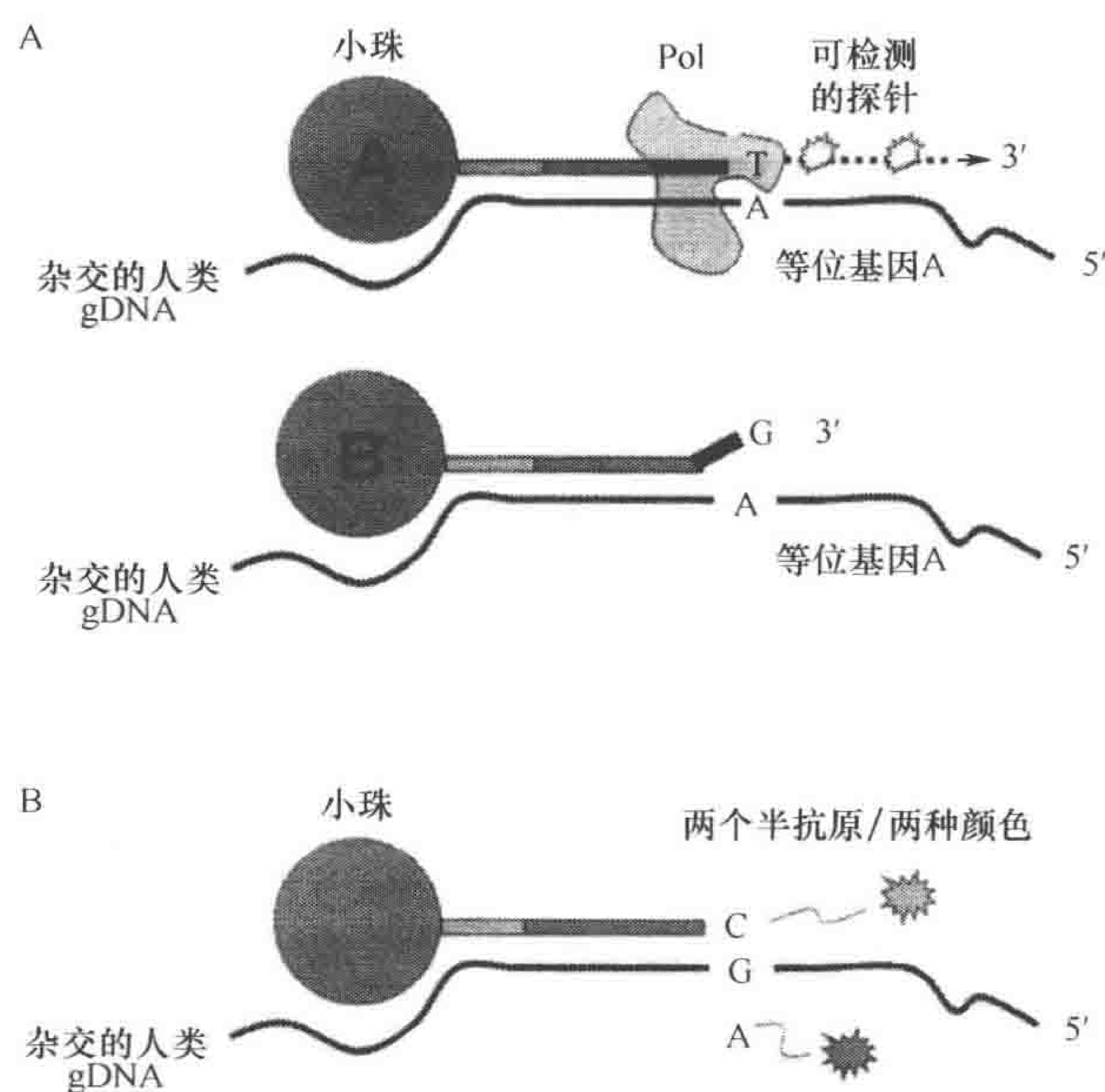



图 13-4 Infinium 检测概况（A）Infinium Ⅰ，等位基因特异性引物延伸（B）Infinium Ⅱ，单碱基延伸反应，见文中详述

对于单碱基延伸（single-base extension，SBE；Infinium Ⅱ）法（图 13-4），在聚合酶和标记寡核苷酸混合物存在下，用染液反复染色。染色完成后，芯片基质用低盐洗脱缓冲液洗脱，包被 XC4，然后在 Illumina BeadChip 检测仪上成像观察。

Illumina 使用的这台检测仪是双色（543nm/643nm）共聚焦荧光扫描仪，像素分辨率为 0.84μm。BeadChips 在信号扩增/等位基因特异性（单色）或单碱基（双色）延伸产物染色过程中掺入荧光素，扫描仪激活 BeadChips 上的荧光素。荧光强度由 Illumina 的软件收集分析。

Illumina 已经提升了产品性能，达到高检出率（>99%）、高精确性（99.5%）水平。这些数据与使用 Illumina 产品进行研究发表的数据一致（Duerr et al. 2006；Helgadottir et al. 2007；Hunter et al. 2007）。

## DNA 分析：拷贝数变异

在基因组水平检测染色体变异，如基因扩增（包括杂合性丢失）或缺失是癌症生物

学和遗传学分析的关键环节。对拷贝数变异（copy number variation，CNV）的研究已经成为群体遗传学和全基因组关联研究的重要组成部分。现在已经知道人类基因组的大量区域包含大规模的 CNV。这些区域被认为有助于形成人群多样性，还可能影响基因表达水平。

在 SNP 含量和检测模式都不改变的情况下，Affymetrix 的 SNP 芯片和 Illumina 的基因分型 BeadChips 都能用来研究 CNV（第 17 章）。从癌症和认知障碍患者采集的样本都能用这些芯片分析。使用 SNP 芯片进行拷贝数分析的一个缺点是探针间距。探针只是拼贴在 SNP 位点存在的地方，使基因组上留下若干段大的空白区。此外，当 SNP 位于基因组发生拷贝数变异的区域，这部分就不能被筛查出来，因此，在商品化芯片上这些区域可能更是未被充分代表的。然而，认识到这个问题后，Affymetrix 和 Illumina 现在都通过增加专门为拷贝数分析设计的探针，增补他们在各自的标准芯片上的 SNP 容量。大体上，这些探针平均分布于整个基因组，调整到均一的解链温度，以保证得到更加稳定的信号。

## 结论

全基因组分析需要巨额的仪器投资以实现自动化。我们没有讨论实验室信息管理系统（laboratory information management systems，LIMS）和进行数据分析所需的生物信息学部分。这方面的信息在本书稿的其他章节涉及一部分（如下一章的第三节），读者将被指引到这些章节。

我们简要介绍了全基因组分析方法，Affymetrix 和 Illumina 这两个公司有产品出售。其他产品发展前景看好，包括其他制造商生产的芯片、流体芯片和全基因组测序。目前，全基因组测序费用太高不能用于全基因组分析。然而，现在的方法正在不断改进最终将促使价格下降，从而使全基因组测序用于染色体变异分析成为可能。

## 参考文献


Altshuler D., Pollara V.J., Cowles C.R., Van Etten W.J., Baldwin J., Linton L., and Lander E.S. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407:** 513–516.

Barrett J.C. and Cardon L.R. 2006. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38:** 659–662.

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316:** 1331–1336.

Duerr R.H., Taylor K.D., Brant S.R., Rioux J.D., Silverberg M.S., Daly M.J., Steinhart A.H., Abraham C., Regueiro M., Griffiths A., et al. 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314:** 1461–1463.

Gunderson K.L., Kruglyak S., Graige M.S., Garcia F., Kermani B.G., Zhao C., Che D., Dickinson T., Wickham E., Bierle J., et al. 2004. Decoding randomly ordered DNA arrays. *Genome Res.* **14:** 870–877.

Helgadottir A., Thorleifsson G., Manolescu A., Gretarsdottir S., Blondal T., Jonasdottir As. Jonasdottir Ad., Sigurdsson A., Baker A., Palsson A., et al. 2007. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316:** 1491–1493.

Hunter D.J., Kraft P., Jacobs K.B., Cox D.G., Yeager M., Hankison S.E., Wacholder S., Wang Z., Welch R., Hutchinson A., et al. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* (in press).

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Kennedy G.C., Matsuzaki H., Dong S., Liu W.M., Huang J., Liu G., Su X., Cao M., Chen W., Zhang J., et al. 2003. Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21:** 1233–1237.

Matsuzaki H., Loi H., Dong S., Tsai Y.Y., Fang J., Law J., Di X., Liu W.M., Yang G., Liu G., et al. 2004. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density

oligonucleotide array. *Genome Res.* **14:** 414–425.
Pe'er I., de Bakker P.I., Maller J., Yelensky R., Altshuler D., and Daly M.J. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.* **38:** 663–667.
Shen R., Fan J.B., Campbell D., Chang W., Chen J., Doucet D., Yeakley J., Bibikova M., Wickham Garcia E., McBride C., et al. 2005. High-throughput SNP genotyping on universal bead arrays. *Mutat. Res.* **573:** 70–82.
Steemers F.J. and Gunderson K.L. 2005. Illumina, Inc. *Pharmacogenomics* **6:** 777–782.
Steemers F.J., Chang W., Lee G., Barker D.L., Shen R., and Gunderson K.L. 2006. Whole-genome genotyping with the single-base extension assay. *Nat. Methods* **3:** 31–33.
Walt D.R. 2002. Imaging optical sensor arrays. *Curr. Opin. Chem. Biol.* **6:** 689–695.

# 14 用于检测 DNA 大片段拷贝数变异的比较基因组杂交

Ilona N. Holcomb and Barbara J. Trask

*Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109; Department of Genome Sciences, University of Washington, Seattle, Washington 98105*



## 简介

在对人类基因组进行全基因组扫描的过程中，发现了 DNA 大片段拷贝数较高水平的变异现象（Iafrate et al. 2004；Sebat et al. 2004；Sharp et al. 2005）。已观察到的拷贝数变异包括插入、删除和重复。单一的变量就能够包括多个基因和含有数百万碱基对的序列。长度大于 1kb，于对照基因组相比具有不同拷贝数的 DNA 片段称为拷贝数变异（CNV）。在目前 SNP 的研究基础上，CNV 出现频率和位置的分类正在研究之中，并且这将揭示存在于大于 1% 人口之中的 CNV 的亚型［Feuk 等将其定义为拷贝数多态性（CNP），2006］。最终，CNV 的基因组目录和它们的基因含量分析将会帮助我们理解这种类型的基因变异对正常表型变异和疾病易感性的影响程度。微阵列比较基因组杂交（array CGH）作为一种强有力的工具，能够对大规模变异提供综合性分析。

CGH 阵列在数千位点进行基因扫描，以得到待测 DNA 和正常参考 DNA 之间的拷贝数差异。如图 14-1 所示 CGH 阵列总览，待测 DNA 和参照 DNA 分别用不同的荧光染料进行标记，两者在目标位点阵列中竞争杂交，每一个目标位点代表一个已知 DNA 序列。目标 DNA 要么被标记，要么在阵列表面上按两倍或三倍被直接合成（通常在载

玻片上)。对每一个目标位点来说，待测 DNA 与参照 DNA 的荧光强度被计算和解释为相对减少（比率小于 1)，相对增加（比率大于 1)，或者相对不变（比率等于 1)。全基因组扫描的优势在于，对于所有目标位点而言，能够在序列草图上找到基因组的精确位置。因此，我们能够了解已知目标位点所代表的基因座的基因含量，也就是染色体上游目标位点与下游目标位点之间的区域。

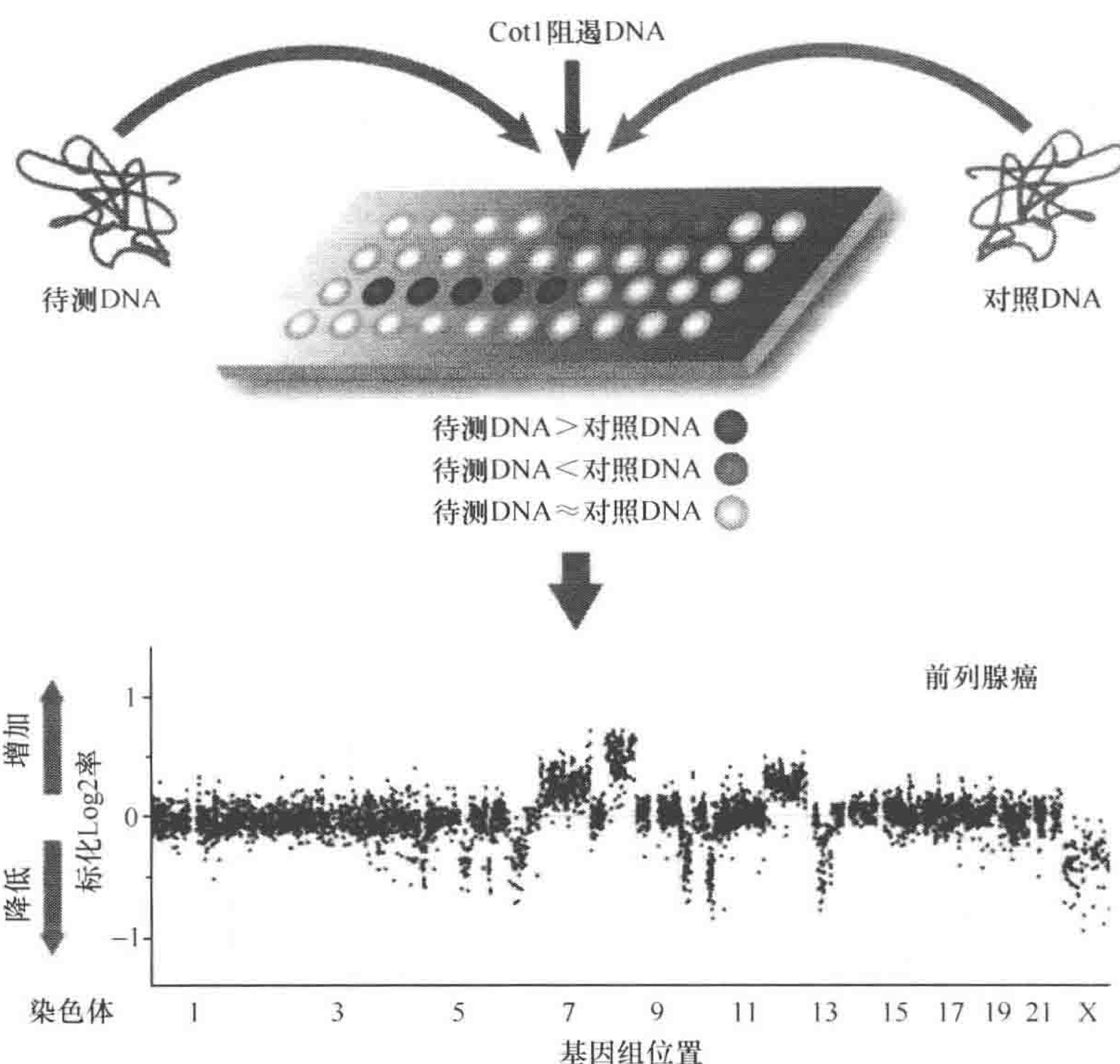



图 14-1 CGH 阵列总览。采用含有共轭荧光染料的 dNTP 对待测 DNA 及对照 DNA 分别进行标记。被标记物经 Cot1 DNA 封闭非特异性重复序列后进行同时杂交。图中所示为前列腺肿瘤细胞拷贝数改变。对待测 DNA 及对照 DNA 的荧光强度比率进行计算和对数转化（以 2 为底)。如图所示，$y$ 轴表示以 2 为底的比率强度的对数，$x$ 轴表示以 Mb 为单位染色体所分开的基因组之间的距离。以 2 为底比率的对数小于 0 则表示待测 DNA 拷贝数相对减少，如大于 0 则表示待测 DNA 拷贝数相对增加，接近 0 则表示拷贝数没有变化

CGH 阵列这个概念开始于作为细胞遗传学经典工具的荧光原位杂交（FISH)。在 FISH 中，由 5～200kb 基因组序列所组成的荧光探针与中期或间期细胞核杂交，来确定序列位置和序列拷贝数。传统 CGH 是 FISH 在全基因组上的应用，它采用了与当今应用于 CGH 阵列相同的原理，但是有一条例外，即传统 CGH 的杂交平台是中期染色体。传统 CGH 的最初发展是用来描述癌症患者病灶癌细胞的总拷贝数的变化（Kallioniemi et al. 1992)。尽管传统 CGH 在当时是一种创新技术，但是它很耗时，并且要求对人类基因组进行精确染色体分类，而且只是粗略地解决了染色体组型异常的界限和位置问题。

第一个 CGH 阵列系统（Solinas-Toldo et al. 1997；Pinkel et al. 1998）的诞生带来了

巨大的进步。这些早期的系统在当时采用了传统 CGH 的竞争法则和由数千基因组克隆所组成的阵列平台。后来发展为具有高产出量和大约 1.5Mb 的分辨率的技术（相对于传统 CGH 20Mb 的分辨率）。最重要的是，应用的每一个克隆采用 PCR 检验的标志序列位点(STS)、末端序列或者它的全序列，都能在人类基因组草图上进行定位。CGH 阵列很快成为癌基因组研究的主要方法，并且越来越多地用于检测先天性疾病的异常。目前，CGH 阵列在描述 CNV 位置、频率和基因含量方面，已逐渐成为一种重要的工具。

如果应用得当，CGH 阵列是 CNV 基因组扫描的有力工具。本章致力于提供一种理解 CGH 阵列基本法则的方法，并且为使用者设计 CGH 阵列试验提供各种可行方案。尤其是下面的 6 个部分分别讨论了不同的实用阵列平台、待测和对照 DNA 的制备、对照 DNA 的选择、杂交要点、数据处理以及当前我们对于人类基因组 CNV 的理解。

## CGH 阵列平台

在理想情况下，应当在人类基因组中以最高的分辨率（就是每 3000Mb 中目标位点的最高数量）和最高的拷贝数变化灵敏度（如既能报告单拷贝变化又能报告多拷贝变化的能力）的条件下应用阵列平台。然而这两个令人期待的特征目前并不能同时存在于同一阵列平台上。讨论基因组分辨率和拷贝数变化的关系，是为了论述这一部分提到的 4 个平台。这 4 个平台分别是：人造细菌染色体（BAC）、寡核苷酸、SNP 和 cDNA 微阵列。由 BAC（100～200kp 的序列）基因组大片段克隆组成的靶位点的阵列，对于基因组变化非常敏感，但是一般基因组分辨率较低（～1.5Mb）(Albertson et al. 2000)。由寡聚核苷酸组成的新一代阵列（100～200bp 序列），分辨率达到数百甚至数千碱基对，但是对于拷贝数差异的敏感性不如 BAC 阵列（Barrett et al. 2004；Carvalho et al. 2004；Sebat et al. 2004）。因此，很有可能牺牲了灵敏度，才获得较高的分辨率。除了这些理论观点之外，在实践上选择平台应考虑的因素还包括阵列分析所要求的待测 DNA 最小量以及每个阵列的相对价值。所有关于 CGH 平台的观点将在下面的段落中阐述。

### 细菌人工染色体微阵列

大基因组克隆（主要是 BAC 克隆）微阵列是第一个 CGH 阵列平台（SolinasToldo et al. 1997；Pinkel et al. 1998）。人们在对数千 BAC 克隆排序和绘制图谱方面所做的大量努力，使得 BAC 克隆成为特定位点拷贝数分析的最佳对象（Cheung et al. 1999，2001；Korenberg et al. 1999；Leversha et al. 1999；Kirsch and Ried 2000；Kirsch et al. 2000）。最初，通过 BAC 末端的序列或 STS，或者是由于被选为人类基因组计划的全序列，数千 BAC 被顺序连接为基因组草图（Cheung et al. 2001；McPherson et al. 2001）。之后，BAC 资源委员会（Cheung et al. 2001）进行的 FISH 作图，有助于区分代表单一定位或多重定位的 BAC（如包含最新复制的序列）。多数 BAC 很容易受 BACPAC 资源中心（http://bacpac.chori.org/）和许多其他发布中心（http://www.ncbi.nlm.nih.gov/genomic/clone/distributors.html)的影响。BAC 克隆包含 100～200kb 的基因组序列，比寡聚核苷酸阵列靶点的平均长度长 10 000 倍。这些大基因组克隆一般比短一些的阵列靶点能产生更

强的信号，理论上产生对拷贝数变化更高的敏感度。通常，基因组上大约在每兆碱基的一个克隆上标记 BAC。这个可粗略地换算为 3000 个阵列靶点。

大量精确作图的 BAC 及其作用，促进了应用于完整基因组和特殊序列的更高密度阵列的产生。先前的一个例子是称为瓦状阵列的新一代 BAC 阵列，包括共同跨越染色体组常染色质部分的重叠 BAC（Ishkanian et al. 2004；Krzywinski et al. 2004；Li et al. 2004）。瓦状阵列在待测基因组高分辨率分析方面是一种非常好的工具，但是完整基因组瓦状阵列要求生产和保持大约 30 000 克隆，这可能会限制它的广泛应用。由单一染色体克隆组成的瓦状阵列更容易生产，在完整基因组典型性分析方面（Woodfine et al. 2004），或者在高分辨率特异性检验导致某种疾病的染色体方面（Ammerlaan et al. 2004），这种瓦状阵列都是非常有效的方法。在检验没有被标准 BAC 阵列很好覆盖的基因组特定序列时，采用富含兴趣区域的定向阵列是另一种有效的方法。应用定向阵列的一个范例是，Sharp 等在研究 CNV 时采用了 SD 阵列（Sharp et al. 2005）。大多数阵列平台在设计时除去 SD，这种 SD 是在基因组多于两个位点之上高序列同一性的（>90%）的较大区域（>1kb）。种种迹象表明，相比较其他基因组位点而言，CNV 在含有 SD 的基因组区域是丰富的（见下面，探索基因组的 CNV；Iafrate et al. 2004；Sebat et al. 2004；Sharp et al. 2005）。因此，相比较包括存在于 SD 之中或 SD 附近区域序列的阵列而言，SD 贫乏序列的阵列可能 CNV 会更少。

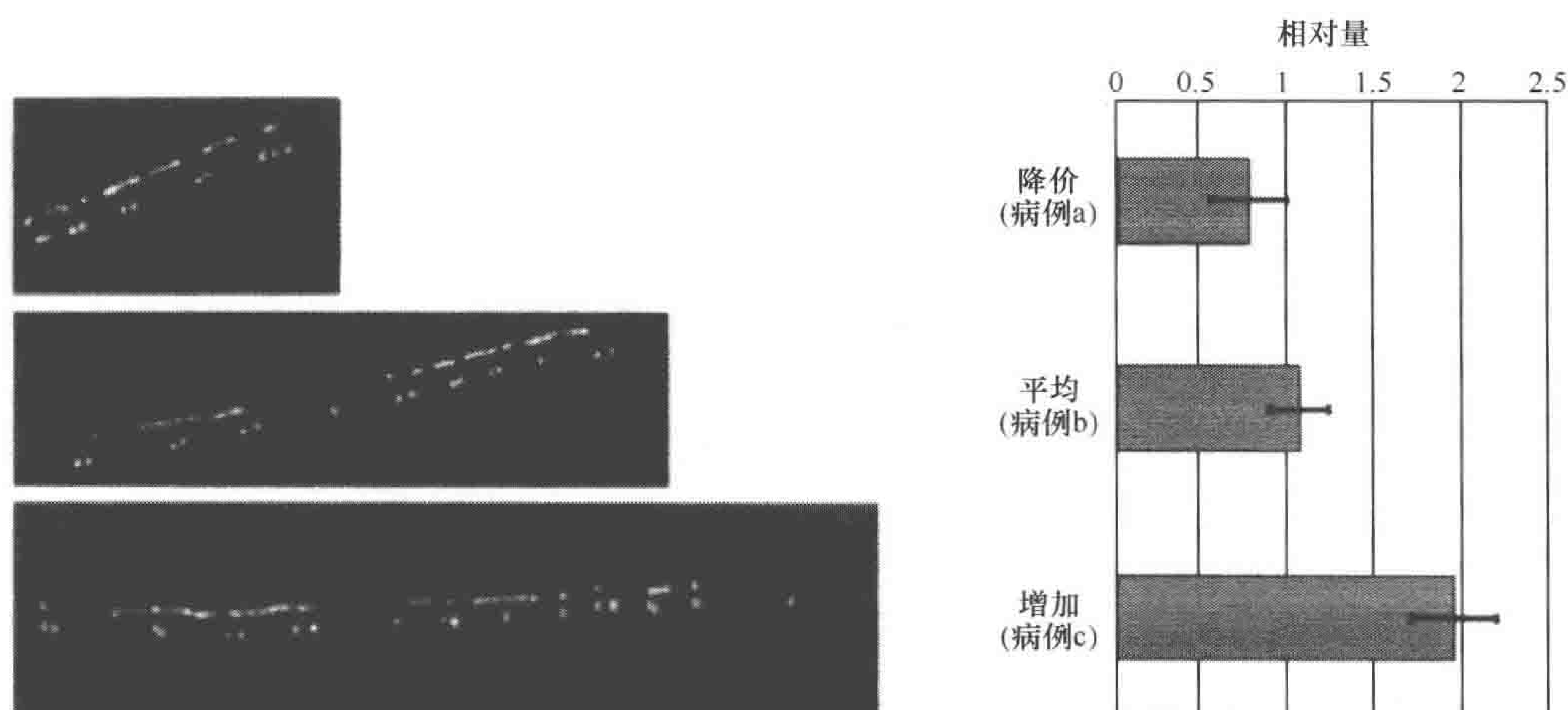



图 14-2　Iafrate 等所做的针对多数一般变异克隆的 FISH 检验在样本人数中的应用。观察到的拷贝数差异反映了克隆中包含的淀粉酶基因衔接重复数的不同。高分辨率纤维 FISH 在延长的 DNA 纤维中被完成，采用了标记绿色荧光素的 RP11-259N12 探针与 5′淀粉酶基因探针（绿色荧光）和 3′淀粉酶基因探针（红色荧光）杂交来检测 CGH 阵列的分布图（没有显示出来的），这种分布图象征由 BAC RP11-259N12 所揭示的 CNV 的相对减少、正常比率或者相对增加。左边的三张照片显示了每种情况的 FISH 结果。相对减少的情况显示了 6 个基因信号（a），正常比率的情况显示了 9 个基因信号（b），相对增加的情况显示了 12 个基因信号（c）。从相同个体 DNA 得到的定量 PCR 结果（如右图所示）与 CGH 阵列以及 FISH 的发现一致（经许可，修改引自 Iafrate et al. 2004 Macmillan）（见图版）

采用 BAC 阵列的三种好处分别是：效益高，比其他平台所需的起始物少，以及能用 FISH 鉴定。第一，BAC 阵列成本低，大约是多数高分辨率寡聚核苷酸基础阵列总设备和处理成本的 1/3。第二，仅从 10ngDNA 得到的扩增产物就能够在 BAC 阵列上得到较好的数据（Guillaud-Bataille et al. 2004；Little et al. 2006）。我们发现，仅从 10 个细胞中取得的扩增产物就已经能够应用于 BAC 阵列，比其他大多数阵列平台所需的起始物少 30 000 倍（I. Holcomb，unpubl）。第三，相同 BAC 报告可疑变化可以用来进行 FISH 鉴定，这是 BAC 阵列的异常特征。如图 14-2 所示，Iafrate 等应用 FISH 分析法确定由 BAC 阵列观察到的拷贝数差异（Iafrate et al. 2004）。在这个表格中，研究中应用 FISH 鉴定三个个体最流行的 CNV 减少、增加和不变（27/55 个体）。和其他研究人员一样，他们也应用定量 PCR 来验证这一发现。

## 寡聚核苷酸阵列

寡聚核苷酸阵列是由短序列（20～100bp）组成的高密度阵列。现已发现在人类基因组近完成序列中的精确位置。这些阵列通常在阵列表面被直接原位合成，这比保持和标记克隆序列节省很多劳动强度。异常高基因组分辨率的阵列的产生，可能是由于较短的序列长度和寡核苷酸生产的简便。当前的寡核苷酸阵列的基因组分辨率是 30～50kb，并且人们正在继续探索生产出更高分辨率的阵列。设计寡核苷酸阵列靶时，通常排除有可能会减少信号干扰比率的序列［如 SD 序列、短散在元件（SINES）、长散在元件（LINES）和 Alu 序列］。然而，SD 的删节可能会减少观察到的 CNV 数量（见前述，BAC 阵列）。不管怎样，阵列靶中减少高度重复性元件，不仅会增加信号与干扰的比率，而且会减少所需的昂贵的重复阻断剂 Cot1 DNA 的数量（参见下述 ，杂交）。为了进一步增加信号和干扰的比率，Lucito 等于 2003 年发明了一项新技术，这项技术被称为代表性寡聚核苷酸微阵列分析（ROMA）。下面的章节将详细论述 ROMA 这种新方法。简言之，ROMA 包括使用 Bg1Ⅱ消化使待测 DNA 与对照 DNA 复杂度下降，以及通过连接介导 PCR 进行的扩增。这种阵列靶序列作为扩增的限制片段的补充，能够提高特异性及增强杂交动力。

目前应用的是短型寡聚核苷酸阵列（～20mer）及长型寡聚核苷酸阵列（～70mer）两种。多数短型寡聚核苷酸阵列是 SNP 阵列。稍长一些的寡聚核苷酸阵列用来优化短序列特异性和较长序列靶阵列敏感性之间的平衡关系。有报道说，长型寡聚核苷酸阵列能够检测出单拷贝变化，并且由 NimbleGen（Madison，Wisconsin）以及 Agilent（Palo Alto，California）开发出可用的商业产品。许多生产寡聚核苷酸阵列的公司（包括那些生产 SNP 芯片的公司），也生产用于标记某物种特定区域的自定义阵列。

## SNP 微阵列

SNP 微阵列是短型寡聚核苷酸阵列的一种，最初用于全基因组基因型分析。如今，SNP 阵列已成为与 CGH 阵列相匹配的平台。像 Affymeteix（Santa clara，Califonia）和 Illumina（San Diego，Califonia）这样的公司已经开发出能够在 SNP 阵列结果中识别拷贝增加和减少的软件。这种在 SNP 阵列上的杂交并非真正的竞争性试验。待测

DNA 得到的杂交强度与对照阵列杂交的平均值相比较，以探明拷贝数差异。SNP 阵列的实验设计以及从中得到的实验数据与其他 CGH 阵列平台是类似的，并且 SNP 芯片的分辨率日益显著增加。Affymetrix（500K 阵列）和 Illumina（HumanHap500 阵列）公司最新产品的基因组分辨率大约为 5kb，是 BAC 阵列平均分辨率的 200 倍。应用 SNP 阵列的另一个好处是这种阵列除了能够报告拷贝数变化之外，还能报告杂合性丢失（LOH）。这一信息对癌症研究通常很重要，因为它帮助区分删除带来的信息丢失（观察到拷贝数减少和 LOH）及节段单亲二倍体（只观察到 LOH）之间的不同。节段单亲二倍体是单亲染色体片段双方拷贝的来源。数百微毫克（nanogram）待测 DNA 的要求以及相比较其他平台的高费用是 SNP 阵列的限制因素。

### cDNA 微阵列

cDNA 阵列曾经长期用于基因表达研究，目前已应用于 CGH 阵列（Pollack et al. 1999）。cDNA 阵列由代表 cDNA 文库中全部或部分基因序列的克隆组成。它们通常包括 10 000～30 000 个阵列靶点。这些阵列的明显优势在于在同一平台上进行 CGH 检测和表达的研究。然而，应用 cDNA 阵列进行 CGH 检测的重要障碍之一，是如何从阵列靶点产生足够信号强度来探测低水平拷贝数变化。一些实验室已经克服了这种局限（Pollack et al. 1999；Chen et al. 2004；Park et al. 2006），但是我们仍期望寡聚核苷酸阵列技术不断进步，以取代 cDNA 阵列。

## DNA 制备

待测 DNA 和对照 DNA 需要用荧光染料分子（通常为 Cy3 和 Cy5）分别标记以用于阵列分析。以待测 DNA 或对照 DNA 为模板，通过线性指数扩增，在序列中参入染料共轭的 dNTP，以完成标记。标记序列的尺度必须符合现行阵列平台的最佳尺度范围。对大多数平台来说，最佳尺度范围为 100bp～1kb，但使用者需要对最佳尺度范围进行检验及相应地对标记方法进行优化。

待测 DNA 和对照 DNA 的品质和数量对阵列数据的质量能够产生深远影响。起始采用降解较少并且大量（>200ng）的 DNA，这是减少数据装置干扰的最有效的方法。符合这些标准的 DNA 也能够用于其他标记方法，直接参入染料共轭 dNTP 的线性扩增是最简单的标记方法，如随机引发和切口平移。随机引发采用 6 碱基顺序简并引物及 Klenow 聚合酶对 DNA 模板进行扩增。这种采用随机引发的有效标记方法要求模板 DNA 完全变性。因此，采用常用的限制性内切酶进行部分降解，对于高相对分子质量 DNA 来说，能够提高反应生产率。随即引发的标记序列大小（100bp～1kb）适合大多数阵列平台。切口平移不需要引物，但是需要脱氧核糖核酸酶在聚合酶起始的 DNA 综合体的模板上制造缺口。在脱氧核糖核酸酶的作用下，切口平移标记序列的尺度范围将会随反应时间的增加而减少。使用者需在琼脂糖凝胶上选择多个时间点检验反应产物，直至找到适当的尺度范围。

严重降解的低品质 DNA，不能用于阵列分析。然而，如果待测 DNA 的数量有限

(<200ng)，样品可能会在应用上述方法之一的标记之前进行一个预先扩增步骤。全基因组 PCR 及多置换扩增法（MDA）能够用于前置放大模板 DNA。两种常用的适合阵列的 PCR 分别为半兼并寡核苷酸引物 PCR（DOP）(Telenius et al. 1992；Daigo et al. 2001）以及连接介导 PCR（LMP）(Guillaud-Bataille et al. 2004)。然而，PCR 扩增的 DNA 一般不能用于商品化的寡核苷酸阵列，因为商品化寡核苷酸阵列通常要求高分子量的待测 DNA。其他实验室选用 MDA，一种相对比较新的非 PCR 基础的扩增方法。据报道，MDA 按照能够应用于 BAC 阵列和寡核苷酸阵列（Paez et al. 2004）的相对较长序列（～10kb）的一段来对基因组进行严格复制。MDA 采用抗核酸外切酶的随机引物以及高持续合成能力的 DNA 聚合酶（如 phi29）对模板 DNA 进行扩增。使用者需要对多种技术进行评定，以确定哪种方法能够产生出最好的阵列数据。

## 对照基因组中的 CNV

采用 CGH 阵列，相对于对照 DNA 对拷贝数变化进行观察。因此，对照基因中的 CNV 会影响待测群体中 CNV 的频率和检验。因此，有必要考虑是采用单个个体的基因组作为对照，还是采用多个个体的基因库作为对照。在单基因组对照中存在的稀有 CNV 是一个潜在问题，因为这种稀有 CNV 会在多数阵列结果中出现，并且会导致对 CNV 出现频率的过高估计。FISH 能够分辨出 CNV 源，或者利用混合对照基因组 DNA 稀释稀有 CNV 的密度能够减轻对 CNV 的过高估计。无法检测待测基因组和对照基因组中共用的 CNV，这对于两种对照类型来说都是一个问题。待测组和单基因对照组中相同的 CNV 的比率为 1，意味着在该基因座没有发生变化。由待测 DNA 及混合对照中的一些个体共享 CNV 的比率值接近 1（不变），随着 CNV 对照个体数量上升。总的来说，随着相同 CNV 对照个体数量的增加以及阵列敏感性的减少，无法探测出 CNV 的可能性正在增加。理想情况下，应当参考相当多的单基因组对照，来检测每一个基因组以准确预测种群频率及优化检测过程。由于符合这些标准的阵列数可能会超出大多数研究的范围，因此需要明确这些对照假设的作用。

## 杂交

通过采取除去未包含在内的核苷的标记反应以及利用复制阻断剂 Cotl1 沉淀标记的 DNA 两项措施，制备待测 DNA 及对照 DNA 用于杂交。Cot1 DNA 是杂交的关键，因为它抑制重复的人类基因组的含量（如 SINES、LINES 以及 Alu 序列），这些重复的基因组含量能够控制杂交，并且强烈减少阵列靶点对拷贝数变化作出回应的能力。一般需要加入 50～100mg Cot1 DNA。

在这里给出的是杂交过程的简要概括，因为通常由阵列设备来完成这一过程。沉淀物样品重悬浮于杂交溶液之中。这种溶液包含甲酰胺、盐类及葡聚糖硫酸酯。如果没有实际情况下较高温度的有害影响，甲酰胺用以增加有效杂交温度，以减少错配。盐类[如含盐的柠檬酸钠（SSC)] 通过降低有效杂交温度，稳定 DNA 混合物。因此，甲酰

胺与盐的适当平衡对于获得有效的杂交（低错配率）是非常必要的。葡聚糖硫酸酯用于增加待测 DNA 与对照 DNA 的有效浓度。杂交通常需要在 37℃的高湿度温箱中温育过夜。下列杂交，错配序列一般通过大约 45℃的甲酰胺与稀盐（如 2×SSC）溶液冲洗载玻片来去除。降低盐浓度，增加甲酰胺的百分率，提高温度都能够增加洗涤的严格性。最后，采用荧光扫描仪收集阵列图像。

## 数据处理

处理阵列输出有三个步骤：在视觉上剔除劣质阵列位点，数据标准化和拷贝数分析。首先，检查出经扫描的阵列图像的缺陷位点（如重叠位点、空白位点或划痕位点）。商业或公众可用的软件（如 GenePix Pro [Molecular Devices，Sunnyvale，California] 及 UCSF Spot and Sproc [Jain et al. 2002]）可用于位点剔除及为下一步处理整理数据。第二步处理工序是标准化，用来在阵列试验中校正系统差异的多重来源。这些差异来自于标记效率的差异，待测 DNA 与对照 DNA 总量的不相等，阵列平台杂交的不均一，以及用于扫描待测 DNA 及对照 DNA 荧光染料输出量的两种激光强度的不同。标准化带来的结果是平均比率或中间比率被设定为标准值。在多数情况下，标准比率为 1，或者对数比率为 1（0）。惯用的标准化程序是：总强度标准化，最低标准化，对数中心化及比率统计（Quackenbush 2002）。自身比对实验有助于找出对已知阵列组来说最好的标准化方案。

鉴别拷贝数变化位点是数据处理的最后一步。最简单的方法是应用基于平均值和标偏差的临界值。然而，有些一流的算法则对于拷贝数的确认是非常有用的。所有这些方法致力于推断拷贝数差异的统计显著性以及确定这些差异的界限。因为象征特殊染色体和基因座位的阵列靶序列共享一个空间关系，这些方法中的多数用于评估邻近序列中每个阵列靶的状况。例如，隐蔽马尔科夫模型（HMM）把相关靶序列空间设定到特定状态。除非这些靶序列的连续亚型的差异比值区分度足以超过设定值的变化，HMM 的情况可以用来说明丢失、获得、或者没有变化。Lai 等（2005）提供了关于 11 种不同方法的探讨，包括 HMMs、混合模型、最大可能性、回归分析和小波动模型。Lai 等人评估的一系列拷贝数工具及算法的可用网站链接是 http://www.chip.org/ppark/ Supplements/Bioinformatics05b.html。

## CNV 基因组探索

从近期发表论文的数量来看，大的相对复杂的遗传变异已经明显地成为正常人类基因组的重要组成部分。在这一领域的三篇开创性论文采用竞争阵列在一组不同民族的个体中筛选鉴定出大量的 CNV（Iafrate et al. 2004；Sebat et al. 2004；Sharp et al. 2005）这里所讨论的一些新发现汇总如表 14-1 所示。

**表 14-1　CNV 全基因组研究总结**

| 研究者 | 分析平台 | 样本量 | 来源 | 总的 CNV | 人均 CNV 值 | 平均大小/kb | 认证的数量 |
|---|---|---|---|---|---|---|---|
| Sebat | ROMA | 20 | individual | 221 | 11.1 | 222 | 11/12 |
| Iafrate | BAC | 55 | pooled | 255 | 12.4 | ～150[a] | 18/18 |
| Sharp | BAC | 47 | individual | 160 | 3.4 | ～150[a] | 7/9 |

a. 修改自 Eichler，2006。

Sebat 等、Sharp 等及 Iafrate 等鉴定的数百个 CNV 表明，我们的研究只是刚开始涉及人类基因组大规模的 CNV。这三项研究所鉴定的全部 CNV 汇编成图 14-3 全基因组图谱。Sebat 等（2004）采用 ROMA 系统评估了 20 个个体，利用代表 76 个单一 CNV 的拷贝数差异来鉴别 221 阵列靶点，平均每人约 11 个 CNV。令人难忘的是，在 Sebat 等鉴定的 76 个 CNV 中，Sharpd 等（2005）就发现了这 76 个位点其中的 53 个。Sharpd 等在研究中应用 SD 标记的 BAC 阵列，在 47 个个体中鉴定出总数达 160 个 CNV（每人 3.4 个 CNV）。Sharp 鉴别出大量新奇的 CVN107，很有可能部分原因是由 SD 富集系统所致（CNV 可能富集于 SD 区域，见下面论述）。然而，令人惊讶的是 Iafrate 等人（2004）采用富集非重复克隆的 BAC 阵列系统，观察到了最大量的 CNV 数量。在 55 个个体中，他们鉴别出 255 个 CNV（每人 12.4 个 CNV）。有趣的是，在 Iafrate 等观察到的 CNV 中有 233 个新奇的。在 Iafrate 的鉴别研究中，仅 11 个分别与 Sebat 和 Sharp 的研究结果相一致。此外，如果忽略变量的趋势（也就是说，丢失或获得），在三个人的研究中，被鉴定的 CNV 仅在 7 个位点上重叠一致（图 14-3）。这 7 个重叠位点中有 5 个 CNV 变化方向相同。三项 CNV 鉴别研究结果缺乏一致性，可能是由于应用了不同的阵列平台，或者由于研究中被测个体 CNV 含量具有有效差异。每一项研究采用独立的方法来确认他们所发现的 CNV 亚型，确认比率比较高。Sharp、Sebat及 Iafrate 小组的确认比率分别是 78 %（7/9）、92 %（11/12）及 100 %（18/18）。

这三项研究所鉴定的全部非重叠 CNV 代表了大约 2.7%（合计 88.35Mb）的单倍体基因组。Sebat 等（2004）采用由 85 000 个寡核苷酸组成的阵列观测到平均长度 465kb 及中值长度 222kb 的 CNV。在这三项研究中，这一阵列的基因组分辨率最高。Sharp 等（2005）及 Iafrate 等（2004）所鉴别的大多数 CNV 包含各自 BAC 阵列上的单克隆。单克隆变量包括它本身和邻近克隆之间的全部或部分插入序列，这使得 CNV 的实际大小更加难以估量。例如，Isfrate 研究工作中所用的 BAC 阵列的克隆之间大约相隔 1Mb，这样增加或减少就涉及 2Mb 的 DNA 序列。然而根据 Iafrate 和 Sharp 的研究结果，遗传变异趋势的检验给出的 CNV 中值大约为 150kb（Eichler 2006）。所有的这些研究发现了一个相等数值 CNV 获得和丢失，及这种变异分布在整个基因组。然而，关于高一致性 SD 区组的 CNV 分布揭示了一种有趣的关系。

这三项研究为 CNV 富集在含 SD 的区域这一观点提供了支持，SD 区域被认为是染色体重排的易发生区域（Bailey et al. 2002）。Sebat 等（2004）在缺失和重复的 CNV 中发现了比完整基因组的平均含量分别高 6 倍和 12 倍的 SD 序列含量。Iafrate 等

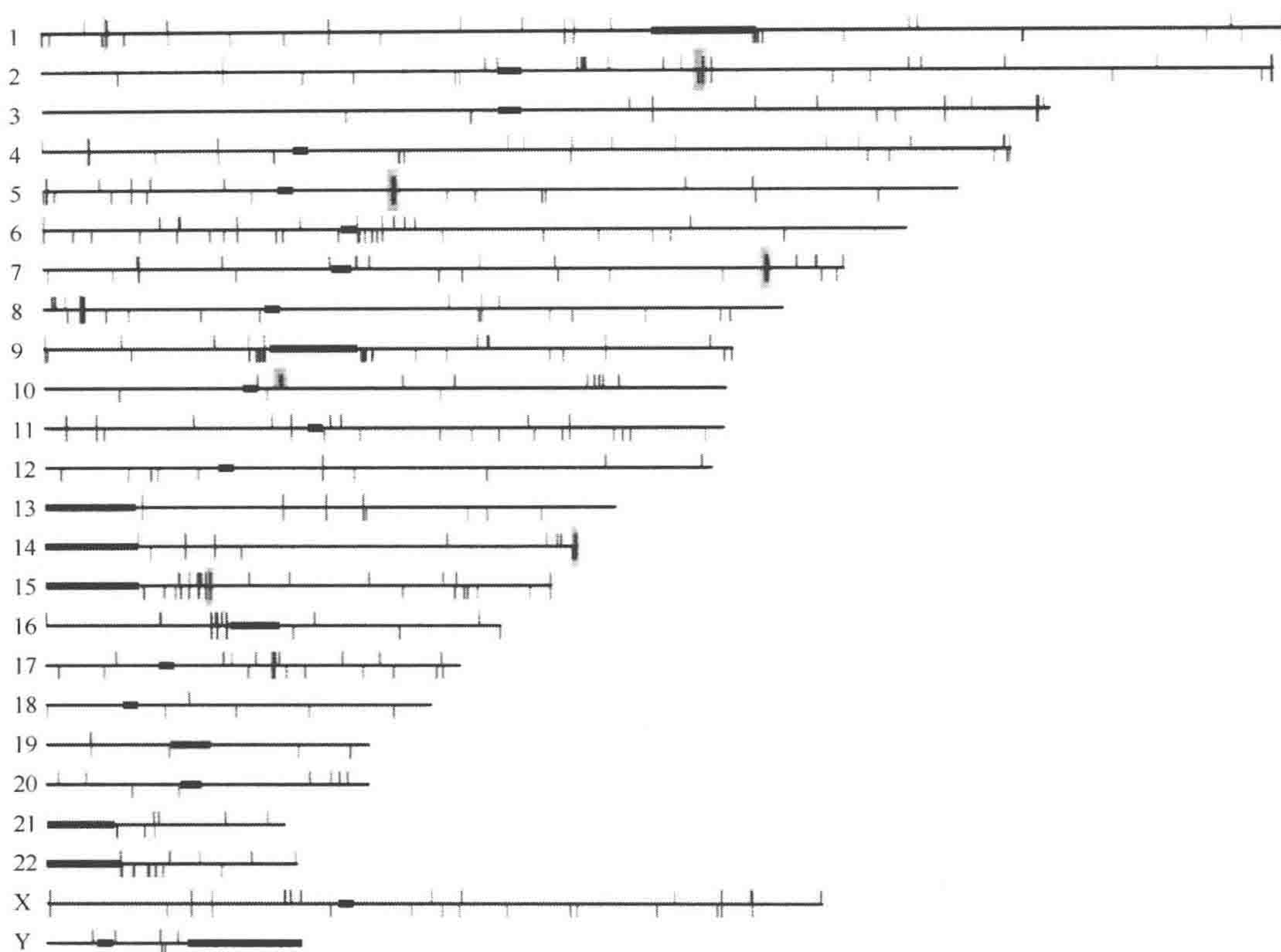



图 14-3　所示 Sebat 等、Lafrate 等和 Sharp 等检测的拷贝数改变位点的全基因组图谱。每一个染色体上的有色条显示了 CNV 的位置（并非按比例显示）。Sebat 等鉴别的 CNV 染色为洋红，Iafrate 等鉴别的为绿色，Sharp 等鉴别的为蓝色。染色体上方的标记符号（并非按比例标记）表示获得，下方的表示丢失。三个人的研究中重叠的 7 个 CNV 位点，与改变类型无关（即：获得或丢失），采用亮灰框突出。这 7 个位点之中的 5 个包含至少与一个研究结果一致 CNV 的变化，这 5 个位点分别在 1、2、5、7、10 号染色体上。基因组在着丝粒和卫星区的集合间隙用暗灰色框显示。图中所示一致的 CNV 来自 2003 年 7 月汇集。可参见网站：http：//humanparalogy. gs. washington. edu/ strcturalvariation/（见图版）

（2004）发现与多重个体中象征 CNV 的大约 25％的 BAC 作对比的区域，是相比较观察到的 7.3％相对于在他们的 BAC 阵列上的所有克隆与 SD 交叠的区域（$p<0.0001$）。Sharp 等（2005）应用 SD 富集的 BAC 阵列，比无 SD 靶点 BAC 阵列鉴定的 CNV 多 11.5 倍。越来越多的研究表明人类基因组的重复结构能够调节正常变异（Bailey et al. 2002；Sharp et al. 2005；de Bustos et al. 2006；Goidts et al. 2006；Redon et al. 2006），SD 与 CNV 之间存在的因果关系的假设引起了人们的兴趣。然而，分辨因果关系需要技巧。通过对基因组富含 SD 区域的彻底调查，可以分辨出含有 SD 的 CNV 的所有的自然的优先结合体。这些区域包括近着丝粒区、亚端粒区及目前 CNV 研究还未勘察的区域。

重要的是，CNV 的亚型包括影响常规表型变异和疾病易感性的基因。总之，Sebat 等总共观察到 70 个基因的拷贝数变异，Sharp 等发现完全或部分与 BAC 变量重叠的 141 个基因，Iafrate 等鉴别出 67 个包含一个或多个完整基因的 BAC 克隆。Sharp 等（2005）发现的 CNV 相关基因包括一些涉及到免疫和新陈代谢的基因，这些对表形变异的形成是很重要的，如原型体抗性和新陈代谢速率等。Iafrate 等（2004）和 Sebat 等（2004）都在正常个体中检测出了包含癌基因的 CNV 的亚型。表面上看，这些基因并不能直接引起疾病，但是结构变异可能会引发染色体重排或者影响基因表达，这会造成肿瘤发生。总之，在 CNV 内部或附近发现的基因可能会对正常生物体或疾病的发生有着深刻的影响。

## 结论

从当前这篇文献的观点来看，CNV 似乎是在人类基因组常规变异中比较突出的部分。总结 CNV 研究成果的两个非常有用的网络资源是 http://humanparalogy. gs. washington. edu/structuralvariation/和 http://projects. tcag. ca/variation/。然而，为 CNV 编纂全部目录尚不成熟。我们仍需要探索精确的尺度范围，合适的基因组序列数量以及 CNV 的基因含量。尽管 CNV 的定义大小为 1kb，但是一些可能包含长达 2Mb DNA 序列（Iafrate et al . 2004）。迄今为止，潜在的 88Mb 的基因组序列存在拷贝数变异（Iafrate et al . 2004），但是进一步的研究显示这一估算数值可能会增加。最终，我们了解到 CNV 可能包含多个基因（Iafrate et al . 2004；Sebat et al . 2004；Sharp et al . 2004），但是我们仍需要确定是否通常的基因包含变异或者 CNV 是否能够影响常规表型变异以及疾病易感性。为了全面评估人类基因组的差异，我们需要在与健康人群全基因组 SNP 谱规模比较的基础上进行 CNV 分析，而 CGH 阵列是完成这一目标的有力工具。

## 参考文献

Albertson D.G., Ylstra B., Segraves R., Collins C., Dairkee S.H., Kowbel D., Kuo W.L., Gray J.W., and Pinkel D. 2000. Quantitative mapping of amplicon structure by array CGH identifies *CYP24* as a candidate oncogene. *Nat. Genet.* **25:** 144–146.

Ammerlaan A.C., de Bustos C., Ararou A., Buckley P.G., Mantripragada K.K., Verstegen M.J., Hulsebos T.J., and Dumanski J.P. 2005. Localization of a putative low-penetrance ependymoma susceptibility locus to 22q11 using a chromosome 22 tiling-path genomic microarray. *Genes Chromosomes Cancer* **43:** 329-338.

Bailey J.A., Gu Z., Clark R.A., Reinert K., Samonte R.V., Schwartz S., Adams M.D., Myers E.W., Li P.W., and Eichler E.E. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003–1007.

Barrett M.T., Scheffer A., Ben-Dor A., Sampas N., Lipson D., Kincaid R., Tsang P., Curry B., Baird K., Meltzer P.S., et al. 2004. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci.* **101:** 17765–17770.

Carvalho B., Ouwerkerk E., Meijer G.A., and Ylstra B. 2004. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J. Clin. Pathol.* **57:** 644–646.

Chen Q.R., Bilke S., Wei J.S., Whiteford C.C., Cenacchi N., Krasnoselsky A.L., Greer B.T., Son C.G., Westermann F., Berthold F., et al. 2004. cDNA array-CGH profiling identifies genomic alterations specific to stage and MYCN-amplification in neuroblastoma. *BMC Genomics* **5:** 70.

Cheung V.G., Dalrymple H.L., Narasimhan S., Watts J., Schuler G., Raap A.K., Morley M., and Bruzel A. 1999. A resource of mapped human bacterial artificial chromosome clones. *Genome Res.* **9:** 989–993.

Cheung V.G., Nowak N., Jang W., Kirsch I.R., Zhao S., Chen X.N., Furey T.S., Kim U.J., Kuo W.L., Olivier M., et al. (BAC Resource Consortium). 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409:** 953–958.

Daigo Y., Chin S.F., Gorringe K.L., Bobrow L.G., Ponder B.A., Pharoah P.D., and Caldas C. 2001. Degenerate oligonucleotide primed-polymerase chain reaction-based array comparative genomic hybridization for extensive amplicon

profiling of breast cancers: A new approach for the molecular analysis of paraffin-embedded cancer tissue. *Am. J. Pathol.* **58:** 1623–1631.

de Bustos C., Diaz de Stahl T., Piotrowski A., Mantripragada K.K., Buckley P.G., Darai E., Hansson C.M., Grigelionis G., Menzel U., and Dumanski J.P. 2006. Analysis of copy number variation in the normal human population within a region containing complex segmental duplications on 22q11 using high-resolution array-CGH. *Genomics* **88:** 152–162.

Devries S., Nyante S., Korkola J., Segraves R., Nakao K., Moore D., Bae H., Wilhelm M., Hwang S., and Waldman F. 2005. Array-based comparative genomic hybridization from formalin-fixed, paraffin-embedded breast tumors. *J. Mol. Diagn.* **7:** 65–71.

Eichler E.E. 2006. Widening the spectrum of human genetic variation. *Nat. Genet.* **38:** 9–11.

Feuk L., Carson A.R., and Scherer S.W. 2006. Structural variation in the human genome. *Nat. Rev. Genet.* **7:** 85–97.

Goidts V., Cooper D.N., Armengol L., Schempp W., Conroy J., Estivill X., Nowak N., Hameister H., and Kehrer-Sawatzki H. 2006. Complex patterns of copy number variation at sites of segmental duplications: An important category of structural variation in the human genome. *Hum. Genet.* **120:** 270–284.

Guillaud-Bataille M., Valent A., Soularue P., Perot C., Inda M.M., Receveur A., Smaili S., Crollius H.R., Benard J., Bernheim A., et al. 2004. Detecting single DNA copy number variations in complex genomes using one nanogram of starting DNA and BAC-array CGH. *Nucleic Acids Res.* **32:** e112.

Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., Scherer S.W., and Lee C. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* **36:** 949–951.

Ishkanian A.S., Malloff C.A., Watson S.K., DeLeeuw R.J., Chi B., Coe B.P., Snijders A., Albertson D.G., Pinkel D., Marra M.A., et al. 2004. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* **36:** 299–303.

Jain A.N., Tokuyasu T.A., Snijders A.M., Segraves R., Albertson D.G., and Pinkel D. 2002. Fully automatic quantification of microarray image data. *Genome Res.* **12:** 325–332.

Kallioniemi A., Kallioniemi O.P., Sudar D., Rutovitz D., Gray J.W., Waldman F., and Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258:** 818–821.

Kirsch I.R. and Ried T. 2000. Integration of cytogenetic data with genome maps and available probes: Present status and future promise. *Semin. Hematol.* **37:** 420–428.

Kirsch I.R., Green E.D., Yonescu R., Strausberg R., Carter N., Bentley D., Leversha M.A., Dunham I., Braden V.V., Hilgenfeld E., et al. 2000. A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nat. Genet.* **24:** 339–340.

Korenberg J.R., Chen X.N., Sun Z., Shi Z.Y., Ma S., Vataru E., Yimlamai D., Weissenbach J.S., Shizuya H., Simon M.I., et al. 1999. Human genome anatomy: BACs integrating the genetic and cytogenetic maps for bridging genome and biomedicine. *Genome Res.* **9:** 994–1001.

Krzywinski M., Bosdet I., Smailus D., Chiu R., Mathewson C., Wye N., Barber S., Brown-John M., Chan S., Chand S., et al. 2004. A set of BAC clones spanning the human genome. *Nucleic Acids Res.* **32:** 3651–3660.

Lai W.R., Johnson M.D., Kucherlapati R., and Park P.J. 2005. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21:** 3763–3770.

Leversha M.A., Dunham I., and Carter N.P. 1999. A molecular cytogenetic clone resource for chromosome 22. *Chromosome Res.***7:** 571-573.

Li J., Jiang T., Mao J.H., Balmain A., Peterson L., Harris C., Rao P.H., Havlak P., Gibbs R., and Cai W.W. 2004. Genomic segmental polymorphisms in inbred mouse strains. *Nat. Genet.* **36:** 952–954.

Little S.E., Vuononvirta R., Reis-Filho J.S., Natrajan R., Iravani M., Fenwick K., Mackay A., Ashworth A., Pritchard-Jones K., and Jones C. 2006. Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA. *Genomics* **87:** 298–306.

Lucito R., Healy J., Alexander J., Reiner A., Esposito D., Chi M., Rodgers L., Brady A., Sebat J., Troge J., et al. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res.* **13:** 2291–2305.

McPherson J.D., Marra M., Hillier L., Waterston R.H., Chinwalla A., Wallis J., Sekhon M., Wylie K., Mardis E.R., Wilson R.K., et al. (International Human Genome Mapping Consortium). 2001. A physical map of the human genome. *Nature* **409:** 934–941.

Paez J.G., Lin M., Beroukhim R., Lee J.C., Zhao X., Richter D.J., Gabriel S., Herman P., Sasaki H., Altshuler D., et al. 2004. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32:** e71.

Park C.H., Jeong H.J., Choi Y.H., Kim S.C., Jeong H.C., Park K.H., Lee G.Y., Kim T.S., Yang S.W., Ahn S.W., et al. 2006. Systematic analysis of cDNA microarray-based CGH. *Int. J. Mol. Med.* **17:** 261–267.

Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W.L., Chen C., Zhai Y., et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20:** 207–211.

Pollack J.R., Perou C.M., Alizadeh A.A., Eisen M.B., Pergamenschikov A., Williams C.F., Jeffrey S.S., Botstein D., and Brown P.O. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23:** 41–46.

Quackenbush J. 2002. Microarray data normalization and transformation. *Nat. Genet.* (suppl.) **32:** 496–501.

Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shapero M.H., Carson A.R., Chen W., et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.

Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Månér S., Massa H., Walker M., Chi M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Sharp A.J., Locke D.P., McGrath S.D., Cheng Z., Bailey J.A., Vallente R.U., Pertz L.M., Clark R.A., Schwartz S., Segraves R., et al. 2005. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77:** 78–88.

Solinas-Toldo S., Lampel S., Stilgenbauer S., Nickolenko J., Benner A., Dohner H., Cremer T., and Lichter P. 1997. Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes, Chromosomes & Cancer* **20:** 399–407.

Telenius H., Carter N.P., Bebb C.E., Nordenskjold M., Ponder B.A., and Tunnacliffe A. 1992. Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13:** 718–725.

Woodfine K., Fiegler H., Beare D.M., Collins J.E., McCann O.T., Young B.D., Debernardi S., Mott R., Dunham I., and Carter N.P. 2004. Replication timing of the human genome. *Hum. Mol. Genet.* **13:** 191–202.

# 互联网资源

**http://bacpac.chori.org/** BACPAC Resources Center (BPRC), Children's Hospital Oakland Research Institute, California.

**http://www.ncbi.nlm.nih.gov/genome/clone/distributors.html** Clone Registry, National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland.

**http://www.chip.org/~ppark/Supplements/Bioinformatics05b.html** Lai et al. 2005. Supplementary material. *Bioinformatics*, 2005. Computational Genomics (PI: Peter J. Park).

**http://humanparalogy.gs.washington.edu/structuralvariation/** Human Structural Variation Database, Department of Genome Sciences, University of Washington, Seattle.

**http://projects.tcag.ca/variation/**

# 15 用以检测遗传变异的展示性寡核苷酸微阵列分析

Rob Lucito

*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724*



## 简介

在历史上，拷贝数目变异（copy number variation，CNV）是通过比较正常和患病个体的核型而首先观察到的。对于某些综合征（从唐氏综合征到癌症）来说，研究核型已成为具有临床意义的重要手段（Benson 1961；Stimson 1976）。一般来说，由于癌症中的 CNV 数目庞大，在特定的区域可以发生广泛的拷贝数变化，对它们的鉴定相当容易。在区域已知的情况下，像 Southern 印迹或荧光原位杂交（fluorescence in situ hybridization，FISH）都可以运用。但是，如果需要对全基因组进行扫描，或者对一个感兴趣的区域没有事先的了解，那么其他技术可能更合适。像比较基因组杂交（comparative genomic habridization，CGH）（Thompson and Gray 1993）那样的早期的 CNV 检测方法很有用，但分辨率太低。随着基因组计划的作图信息和微阵列技术的发展，CGH 在阵列分析平台中派上了用场（Pinkel et al. 1998；Pollack et al. 1999），人类基因组序列测定的完成为具有精细分辨力的阵列 CGH 方法的发展铺就了道路（Lucito et al. 2003；Barrett et al. 2004；Selzer et al. 2005）。随着增加的分辨率，可以很明显地观察到，在正常群体中 CNV 出现频率很高但先前却未曾检测到（Lucito et al. 2003；Sebat et al. 2004）。

在 Michael Wigler 博士的合作下，我们发展了基于阵列的 CGH 方法用来检测癌症中的 CNV，尽管我们很快就认识到它在确定正常种群的小 CNV 中非常精确。这种称为寡核苷酸微阵列展示分析（ROMA）（Lucito et al. 2003），是展示性的差异显示分析（RDA）——一种由 Wigler 博士和 Nikolai Lisitsyn（Lisitsyn et al. 1993）发展的技术。两种方法的核心都是进行基因组展示的实践。一个展示过程就是一个可再生的基因组抽

样分析。它的制备程序是，以一种限制性内切核酸酶对基因组进行裂解、用 PCR 接头进行连接和用一种适合于 PCR 的聚合酶（如 Taq 聚合酶）进行扩增。一般来说，当模板是混合大小的片段群体时，一种像 Taq 这样的酶将优先扩增较小的片段，导致大小为 100～1200 bp的展示结果。较大片段的扩增缺陷导致样本复杂性的减少。相对于全基因组杂交而言，展示结果复杂性的减小使得信噪比有所增加，这要归功于增加了的杂交效率和片段标记效率。片段标记效率的增加是阵列与展示内容高度匹配的一个结果。换句话说，阵列的核苷酸内容比全基因组更加紧密地与所展示的内容相匹配。

### 展示法

展示的复杂性可以由用于基因组 DNA 起始消化的限制性内切核酸酶来定型。例如，当一个具有六核苷酸切点的限制性内切核酸酶被使用时，复杂性降低到原来基因组的 2%～5%；当一个具有四核苷酸切点的限制性内切核酸酶被使用时，复杂性降低到原来基因组的 60%～80%，还是依赖于内切核酸酶的裂解位点。展示的复杂性还可以由展示的损耗来进一步定型。在这一过程中，第二种限制性内切核酸酶在 PCR 扩增之前裂解开基因组 DNA 的连接。这使得我们可以随意调整实验复杂度。

作为所需要的 PCR 扩增的一个结果，展示体系可以从少到 50 ng 的基因组 DNA 中很容易地制备。这使得人们很容易地分析不能恢复的样品（如肿瘤标本），其中产量常常在接近100 ng的范围。超量扩增可能发生：随着极端数目的循环，测量的比率下降。在大多数情形下，这不是什么问题。然而，半合子或重复的小 CNV（包含阵列上少数探针的区域）的检测却会产生很多问题。我们对随循环数目变化的样本制备的展示体系进行比较，发现进行多到 25 个循环的体系中，在 CNV 的检测水平上并无任何变化。目前，在上述讨论的模板条件下，我们不推荐多于 25 个循环的 PCR 扩增。

## 影响 ROMA 的因素

有几种因素影响在特定群体中对 CNV 进行检测的能力，包括阵列分析本身、展示、物种类型和试验的设计问题。展示分析在反映与参照样本相关的基因组拷贝数方面是相当精确的。为了精确地检测一个基因组中的 CNV，所有能够控制住的系统干扰成分源都必须封堵。通过 PCR 进行的样本的扩增会引入可能的干扰源，这很容易被关注。应当使用数量可比较的基因组模板 DNA，并且从限制性内切核酸酶消化到 PCR 扩增，实验样本和对照样本都应该平行制备。理论上讲，也可能对某一单个样本进行 ROMA 而不使用对照，但目前这种方式还在研究中。

当一个研究者想要在一个普通样本中检测 CNV 时，阵列分析的设计是同等重要的。如果去除这些区域，就只能检测到非常少量的变异。当然，要在基因组中增加不止一次被发现的区域，是必须小心处理的，就像任何基于杂交的手段那样，一种靶位在基因组中出现的次数越多，杂交的特异性就越小。对比一下就可以知道，由于重复区域在基因组中太多了，它们在用基于杂交的方法进行拷贝数检测的研究中没有什么用处。

在阵列上用探针所进行的检测应该大体上等同。倘若探针分析是零星分布的，那么

一种被若干种探针（在那里内部探针是很差的报告因子）所覆盖的 CNV 就会被解释为一簇 SNP。在这种情况下，CNV 检测和（更加特异的）自动化检测形式容易漏掉这样的 CNV。在阵列分析中应该赋予相似方式运用的探针，或者，对探针利用参数的使用应该能够对输出数据进行解释。现在已经发展出预测探针运用方式的方法，而且现在的 ROMA 阵列分析都采用相似利用方式的探针。

在阵列分析中，分辨率也同等重要。如果对于一个给定的区域没有足够的探针，CNV 检测就无从谈起。许多区域是相当小的（100～200 kb），而探针数少于40 000个的阵列分析将不能检测这些区域中的许多成员。探针的替换可能也是一个问题。对于一个细胞来说存在着很关键的未翻译 RNA，它们可能以癌基因的方式发挥作用。因此，基因中心（gene-centric）CGH 阵列分析并不是一种最理想的设计。起初的ROMA阵列分析大体上有 82 000 个探针（虽然可以达到更高密度的水平），它们随机分布于整个基因组，因而所有区域（除高度重复的区域外）在阵列分析中得以同等表现。

在大多数情况下，当 ROMA 被用于鉴定普通样本的遗传变异时，对照样本将从一个不同的个体中取得。既然展示分析是利用限制性内切核酸酶对复杂性进行裁定的，此方法就揭示了限制性片段长度多态性（RFLP）。如果在一个实验样本中存在一种 SNP 而起初的限制性酶切位点遭到了破坏，那么，与对照样本相比就产生了一种 RFLP。通常，新的片段因为太大而难以有效扩增，因而就在展示分析中不再出现。识别这种片段的探针将会在对照样本中取得很高强度的度量值，而在实验样本中度量值很低，这就产生了很高的比率，使这些探针能够从大多数探针测量中分散出来。当这些探针很多时，就可能转移人们的视线，但是数据的质量不会受到影响。实际上，既然这些测量是 SNP 检测的结果，它们就可以在利用分段运算法则（segmentation algorithm）进行 CNV 信息学鉴定的分析中作为有利条件，见下文的讨论。

SNP 的检测在实验生物（如小鼠）的分析中某种程度上被夸大了。相对于相当杂合的人类基因组来说，实验小鼠是极其纯合的，除非近期发生了可假定的生殖细胞遗传转变。小鼠 ROMA 阵列分析是基于可获得的第一份小鼠基因组序列资料的，这一资料几乎全部来自 C57BL6/J。所以，如果以 C57BL6/J 作为参比进行任何其他物种的阵列分析，大量的 SNP 事件将被检测到。这些事件跨越基因组聚集成片状的分布，因此在某些区域的密集位点一眼就能辨别到大的 CNV 成簇出现。有趣的是，与 C57BL6/J 相比，SNP 簇的基因组位置从一个品系到另一个品系是有变异的。不同品系间的变异意味着试验的设计必须非常谨慎，因为错误的参照样本会造成无法解释的结果。可能的话，最简单的解决办法是利用来自同样小鼠的材料。虽然家族间略有遗传差异，但家族内的选择是第二种最好的选择（I. Hall 未发表）。如果要编制不同品系的目录，利用特定小鼠的基因组架构框架是可以描绘其血统关系的。对生殖细胞和体细胞的 DNA 进行比较，可以使人们在许多世代之间鉴定生殖细胞 DNA 在内切核酸酶位点上的改变。这进而也使人们能够确定动物中的突变率，鉴定任何可能的突变热点。

当人类 ROMA 被用于分析正常拷贝数目的变异时，参照样本往往来自不同的个体，除非比较的是体细胞 DNA 和生殖细胞 DNA。所检测的所有 CNV 必须与参照样本进行对比来诠释（在 SNP 分析中这一般不是问题）。因此，最好尽可能使用很少的参照

样本DNA，可能的话一雄一雌为好。利用同样的参照样本使得对一个特异CNV的拷贝数的鉴定变得很容易。另外，无论在何时可能的情况下，分析亲代以追踪CNV的遗传问题都是有用的。当要寻求最初的改变时，这更是基本的要求。

## 阵列分析的探针设计

在阵列上使用的所有探针被设计得与展示片段互补，也就是说，与小的限制性片段互补。探针设计开始于基于可获得的基因组序列所进行的芯片展示（silico representation）。通过对感兴趣的限制性位点的所有例子进行定位和记录，并选择100～200 bp的大小，这一分析可以很容易地进行。相分离地进行分析，整个基因组被划分为不同长度的N-mer单位，最普遍的是15个和21个核苷酸长度。作为探针选择的前体，所有被鉴定的展示片段作为全部可能重叠的50-mer大小（阵列分析的最终探针长度），与基因组的N-mer文库进行比较，并用表示在N-mer的基因组文库中发现一个N-mer片段的次数的N-mer频率进行解释（更详细内容见Healy et al. 2003）。这种分析所涉及的是长度为N=21和N=15的N-mer单位，利用的是Burrows-Wheeler转化法（Burrows and Wheeler 1994）。对于处于芯片内消化（in silico digest）之中的每一个片段，我们对每一个可能的50-mer单位进行定量特性分析。21-mer频率被用作基因组中整体唯一性的度量，而我们的更严格标准是所有有候选资格的21-mer必须是唯一的。不过，这在分析正常的变异时又过分严格了。对于那些已经通过21-mer唯一性衡量的探针来说，15-mer频率被减到最小，因为它们的作用是预测量度交叉杂交的可能性。

探针的选择是按照总体N-mer频率的下行顺序进行的，本质上是一个过滤程序。这将候选探针原本无边的搜索空间限制到那些很可能给予我们准确的比率量度而面临交叉杂交事件或造成合成过程失败的可能性尽可能小的范围。然后，完全根据经验进行探针测试。很典型的开始步骤是设计和分析比适于一次阵列分析的探针数量多10倍的探针。探针以某种程度的重叠在多个阵列上进行阵列分析，以实现比较测量。每一个阵列都在一种酶去除（enzyme-deletion）实验中使用，其中一个展示分析与一个已经进一步以另一种酶消化处理的等同的展示分析进行比较。既然基因组的完整序列已知，我们可以精确地预测哪一些片段应该在第二步消化中被减除。我们选择那些在非消化途径中具有最高信号强度的探针，因为我们已经确认这是进行实验的好方略。探针在阵列表面上的放置应当是随机的。在这种方法中，阵列表面不出现人为的杂交假象，将呈现出一致的基因组改变信息。

## 杂交后试验分析

在杂交之后，阵列在一个标准的扫描装置上进行扫描。这种装置的一个品牌是GenePix 4000B。然后，可以用许多软件包对信号强度进行采集。GenePix扫描装置具有的最大分辨率是5μm。这对于平衡像素强度来说很重要，平均小于9个像素的情况是不可取的。在最大分辨率情况下，小于15μm×15μm的图像特征是难以分析的。现在使

用的阵列分析具有 17μm×17μm 的图像特征。在分析中，Axon 扫描装置也是适用的。对于使用较小图像特征的高分辨率的阵列分析，就必须采用别的分析软件了。数据采集以后，阵列的分析度量就可以通过好几种不同的方法进行标准化了。现在我们使用的是一种 Lowess 曲线，适合于自 Terry Speed 博士及其同事描述的方法改进的运算法则 (Yang et al. 2002)；接着是一种运算，用于矫正局部的杂交假象。在相邻的探针之间，度量中的拷贝数数据具有变异，源于不同的因素，其中包括阵列杂交中的干扰信号（“噪声”信号）、标记效率、洗涤情况和展示程序。一些这样的变异可以通过重复实验来消除。对于在正常样本中评估变异相对小（以 bp 来计）的遗传变异并且变化常常只是由于一个拷贝的改变所造成情况来说，一次重复实验（如果不能再多的话）是强烈推荐的。杂交可以简单地进行重复，但对于双色杂交（如利用 cy3 和 cy5），最好进行一个色彩逆转（或交换染色）试验，其中样品采用与第一次杂交对换了的染色标记。标准化以后，探针比率就可以进行平衡了，或者可以采用其他方法进行干扰信号的消除。一个例子是单色杂交引起的探针外露现象。通过对比重复实验，这些外露信号可以扣除。对于特异的图像特征，如果在一个实验中的比率高于一个标准偏差，就要采用低比率的探针。这可能去掉了一些信息内容，但是我们发现，一般来说实验数据的质量得到了提高。

即使进行了实验的平衡，数据还是会有固有的干扰成分。可以通过许多不同的运算方法来解除这样的干扰。最简单的方法是移动平衡（moving average）(Pollack et al. 1999)，在基因组顺序中进行移动而对邻近探针进行平衡，被平衡的探针数目可以调整。这改进了增加平滑度的程序，但降低了测量的分辨率，因为通过断裂点的探针的平衡降低了在断裂点末端的探针的平均值。Pollack 等通过进行常态对常态（normal-normal）的对比改进了此方法，在对比中使得一个阈值的计算能够揭示虚假现象。这一方法在癌症拷贝数目检测中相当重要，但对于常规变异的度量来说具有局限性，因为小的损伤可能被错过。其他方法更加棘手，并且要对数据进行模型分析（Hodgson et al. 2001; Autio et al. 2003; Snijders et al. 2003)。我们已经使用了一种由 Olshen 等（2004）开发的运算方法和另一种由 Bud Mishra 开发的方法，用于 ROMA 数据，而它们也可以很容易地用于其他 CGH 数据（Daruwala et al. 2004)。第三种运算方法由 Lakshmi Muthuswamy 和 Micheal Wigler 在冷泉港实验室开发，明确地适用于 ROMA 数据分析，也可以用来进行其他 CGH 平台的分析（Lakshmi et al. 2006)。这三种方法的每一种都有模型基础，其中 Olshen 及其同事开发的方法基于二元分段法（binary segmentation）但使用了数据的参比分布来计算系统干扰信号；Mishra 及其同事利用一种 Bayesian 模型，这种模型可以看作一种能使评价功能最小化的最优化程序。这两种模型都能够精确地鉴定癌症中的拷贝数目改变，而不仅常规的遗传变异，但是相对于由 Muthuswamy 和 Wigler 开发的方法，它们需要明显多的运算时间来分析实验。

由 Muthuswamy 和 Wigler 开发的方法是一种折中的方法，它将强度值的 log 比率分解为均等的部分。比率被安排在基因组次序中，并被以任意分界线分为 100 个数据点 (data points）的模块。然后，通过将方差最小化而使分界线交互式地移动，并进一步通过运用一种 Komogrov-Smirnov（KS）二项分布进行无效假设检验（null hypothesis

test)（Conover 1998），进行精确判定。只有给出一个小于 $10^{-5}$ 的 $p$ 值的分界线才可以接受。由于 KS 检验的本质属性，小于 3 个探针长度的片段一律不予考虑。对于那些包含 3 个或更多探针的损伤来说，用这种方法分析非常好。一种更为精确的分段分析方法正在开发，利用的是隐性 Markov 模型（a hidden Markov model）。这种方法利用 SNP 的存在去确定两个基因组之间的变化状态，也就是说 0 拷贝、1 拷贝、3 拷贝或更多个拷贝。一旦这个指标被确定，其他改变的状态就可以根据起始的运算得以鉴定了。

经过处理，数据就变得可见了，用来鉴定存在于常规样本和参照样本之间的 CNV 和任何 SNP。既然鉴定的改变是与参照样本比较出来的结果，就可能有一个 CNV 在测试样本和参照样本中都以 3 个拷贝形式出现的情况，因而差异性被取消。这对于稀有的 CNV 是不可能的，但是对于普遍的 CNV 则很有可能。如果一个样本具有一个比在另一个样本中发现的略微大一些的 CNV，则结果将显示为两个 CNV，因为中间的区域可能被取消。当对数据进行分析时，这种可能性必须考虑进去。

用其他方法，可以进行大型变异的确认。很明显，用来确定 CNV 变异的最精确的方法是荧光原位杂交（fluorescent in situ hybridization，FISH）。这一方法可以用来估计一个样本中一种等位基因的真实拷贝数目，而 CGH 数据则只能用来鉴别与参照样本相比较的水平。然而，FISH 在通量上来得慢而且需要完整无缺的细胞。另一种方法是定量 PCR（qPCR），此法无需全细胞样品，可以在中等通量水平进行，并且只需要少量的 DNA 模板。我们已经使用这种方法对在大型 ROMA 数据组中鉴定出的 CNV 的相对拷贝数进行常规的确定。拷贝数的计算总是要与一个参照样本进行比较的。如果已经知道参照样本中包含了所研究区域的两个拷贝，则测试样本中真正的拷贝数就可以精确地确定了。最精确的确定需要另一个探针，其拷贝数不仅对于测试样本而且对于参照样本都是已知的。

## 大型样本组的选择和分析

虽然我们可以确定一个在单一样本中鉴定出来的 CNV，但拷贝数的检测还要经常用于对大型样本组进行分析。这可以在常规个体的一个大型样本组中进行，用来确定人类（或小鼠）常规种群中的变异。另外，你也可以对一个患有某种综合征或疾病的大型病例或疑似病例个体群组进行分析，其鉴定得的 CNV 可以结合起来用于确定是不是某个研究对象与特定的综合征或疾病相关。对于所分析的样本来说，个体的某些历史背景需要清楚。例如，当分析胰腺癌的家族性易患性时，就有必要知道所研究的个体以及在紧接着的家族中至少有一个得了胰腺癌。因为胰腺癌在全世界范围所累积的个体数目相当小，在同一个家族中若具有两个患病个体，就增加了癌症有家族性特征成分的可能性(假定环境因素可以排除)。对于疾病或综合征的研究来说，越是纯合就越好。如果一种综合征包含了两种以上真正的疾病，对于任何一个综合征候选区域的确定都更加困难。即使对于家族性易患型的胰腺癌来说，也可能涉及不止一个基因。对于产生一套有价值的 CNV 来说，仔细地选择测试样本组是势在必行的。由于原代细胞不能作为 DNA 的来源，样本的细胞常常用 EBV 感染以使其转化，以获得大量的生长体系。这样的材料

应当谨慎处置，因为转化过程和培养中时间阶段的延伸本身都会导致基因组的异常。

## 大数据组的分析

为了确定哪些被鉴定的 CNV 与所研究的疾病有潜在联系，在数据体系中排除或滤掉越多的正常 CNV 越好。你需要收集一套正常的 CNV 制作一个 CNV 目录，用这个目录将所测试的数据体系中的正常 CNV 排除。一个常规 CNV 的确定在某种程度上是任意的，取决于测试的那套对象，也就是说，常规数据体系随着所研究的综合征的不同是会发生变化的。例如，如果测试组是关于一种或多种癌症的易患人群的，那么目录中所有来自癌症易患家庭的正常数据都应该剔除。这样，常规数据目录中个体的精确家族数据是必不可少的。对于比较一个带有罕见癌症（如胰腺癌）的样本组目录的选择是相对容易的。有严格的法则用于剔除已经具有胰腺癌或已经有直接的家族成员患有胰腺癌的“正常个体”数据。然而，对于有很高的零星发生率或具有低外显率综合征的普通癌症来说，选择较为困难。另外，一个目录成员可能是一个无症状的受影响个体。如果这样的一个个体进入正常的数据组中，在过滤过程中有意义的区域将可能丢失。下面我们将陈述一种拯救这样数据的可能途径。

为 CNV 过滤而准备的常规实验组的大小应当最大可能地使得测试样本组中的普通 CNV 能够剔除。当然，无论过滤体系变得多大，都会有一些罕见的 CNV 不能从群体中过滤掉。如果正常样本组太小，测试数据组就会包含相当大量的不同正常 CNV。一旦正常样本组的选择完全，ROMA 就可以进行了。另外，在公共条件下可以获得的其他实验室的正常数据组也可以用来过滤普通的 CNV。但这只有当准确的患者信息可以取得的时候才可以做。

目前的 ROMA 阵列分析具有 85 000 个探针，但是我们已经开始进行使用390 000个探针的阵列分析。在很多例子中，不仅测试样本的数据组，就是正常样本过滤数据组也可以非常大。如果是基于所有探针的，那么数据组就能获得一个大型数据框架用以展示所有样本，并可以缩短运算所需要的时间。来自基因组的许多数据并不报告什么事件，或者报告的是没有 CNV 的辽阔区域。通过只对鉴定的 CNV 进行展示，数据可以极显著地减小而不至于削弱信息内容。一旦这样的转化在所有样本中进行，能在过滤测试样本中运用的正常组的一个主要档案就产生了。

为了过滤，在测试组中发现的正常组的所有 CNV 都要被减除。当在测试组中发现的一个 CNV 比在正常过滤组中发现的 CNV 在长度上更大（被更多的探针覆盖）时，CNV 的片段常常成为剩余的残留物。在最差的例子中，与发现于正常组的 CNV 相比，测试组中的 CNV 在两端都是较大的。在过滤之后，将出现很接近的位置有两个 CNV 的情况。整个 CNV 都可以被去掉，但是这必须做得非常小心，因为有可能丢失有价值的数据。在这一过程的最后，一个 CNV 在测试组群体发现的频率就能够计算了。将在正常群体中发现的 CNV 剔除，是一个严格控制的操作，特别是存在一个无症状个体的非特异包含物时。一种备选的剔除正常群体中发现的区域的策略是掌握检测水平的机动性。如果一个 CNV 在测试组被发现，而只存在于一个正常组个体，它是不被剔除的。

这种检测水平或机动性可以增加到被确信对于数据组有信息学价值的任何高度。这样的机动性的缺点是它既会增加某些普通的 CNV，也会增加碰巧两种数据组中都存在的稀有 CNV。所以，通过几种不同标准分析被过滤的数据是很有用的。

一旦在测试组中的 CNV 区域被鉴定，其他分析就可以开始了。作为一个例子，区域的协调可以与利用别的技术进行的其他作图手段［如 SNP 作图和数量性状座位（QTL）分析］的结果进行比较。信息学分析可以用来鉴定区域中的所有基因以确定是否有任何基于基因功能和所研究的症状或疾病的有趣的候选基因。要确定所有 CNV 是不可能的，因为这一过程是耗时耗力的。然而，用另一种方法验证一定数量的 CNV（可能是那些特异的感兴趣对象）是很重要的。FISH 是最精确的验证方法，但是 qPCR 是一种合理的备选方案。它具有某种难度，并且需要大量的复制本，因为预期差异可能只有一个循环。虽然如此，如果不能获得完整的细胞，它还是一个可行的选择。

# 参考文献


Autio R., Hautaniemi S., Kauraniemi P., Yli-Harja O., Astola J., Wolf M., and Kallioniemi A. 2003. CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics* **19:** 1714–1715.

Barrett M.T., Scheffer A., Ben-Dor A., Sampas N., Lipson D., Kincaid R., Tsang P., Curry B., Baird K., Meltzer P.S., et al. 2004. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci.* **101:** 17765–17770.

Benson E.S. 1961. Leukemia and the Philadelphia chromosome. *Postgrad. Med.* **30:** A22–A28.

Burrows M. and Wheeler D.J. 1994. A block sorting lossless data compression algorithm. HP technical report SRC-RR-124, Hewlett-Packard, Palo Alto, California.

Conover W.J. 1980. *Practical nonparametric statistics*, 2nd edition. Wiley, New York.

Daruwala R.S., Rudra A., Ostrer H., Lucito R., Wigler M., and Mishra B. 2004. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl. Acad. Sci.* **101:** 16292–16297.

Healy J., Thomas E.E., Schwartz J.T., and Wigler M. 2003. Annotating large genomes with exact word matches. *Genome Res.* **13:** 2306–2315.

Hodgson G., Hager J.H., Volik S., Hariono S., Wernick M., Moore D., Nowak N., Albertson D.G., Pinkel D., Collins C., et al. 2001. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat. Genet.* **29:** 459–464.

Lakshmi B., Hall I.M., Egan C., Alexander J., Leotta A., Healy J., Zender L., Spector M.S., Xue W., Lowe S.W., et al. 2006. Mouse genomic representational oligonucleotide microarray analysis: Detection of copy number variations in normal and tumor specimens. *Proc. Natl. Acad. Sci.* **103:** 11234–11239.

Lisitsyn N., Lisitsyn N., and Wigler M. 1993. Cloning the differences between two complex genomes. *Science* **259:** 946–951.

Lucito R., Healy J., Alexander J., Reiner A., Esposito D., Chi M., Rodgers L., Brady A., Sebat J., Troge J., et al. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res.* **13:** 2291–2305.

Olshen A.B., Venkatraman E.S., Lucito R., and Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5:** 557–572.

Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W.L., Chen C., Zhai Y., et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20:** 207–211.

Pollack J.R., Perou C.M., Alizadeh A.A., Eisen M.B., Pergamenschikov A., Williams C.F., Jeffrey S.S., Botstein D., and Brown P.O. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23:** 41–46.

Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Maner S., Massa H., Walker M., Chi M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Selzer R.R., Richmond T.A., Pofahl N.J., Green R.D., Eis P.S., Nair P., Brothman A.R., and Stallings R.L. 2005. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer* **44:** 305–319.

Snijders A.M., Pinkel D., and Albertson D.G. 2003. Current status and future prospects of array-based comparative genomic hybridisation. *Brief Funct. Genomics Proteomics* **2:** 37–45.

Stimson C.W. 1967. Possible causes of mongolism (Down's syndrome) and other chromosomal aneuploidies. *Mich. Med.* **66:** 436–441.

Thompson C.T. and Gray J.W. 1993. Cytogenetic profiling using fluorescence in situ hybridization (FISH) and comparative genomic hybridization (CGH). *J. Cell. Biochem. Suppl.* **17G:** 139–143.

Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., and Speed T.P. 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30:** e15.

# 16 FFPE 样本拷贝数变化的检测——全基因组取样分析法

Sharoni Jacobs

*Affymetrix, Santa Clara, California 95051*



## 简介

### 全基因组样本分析（WGSA）和拷贝数检测

人类群体和个体在核苷酸序列和 DNA 片段拷贝数（CN）这两个方面都有广泛差异。序列分歧普遍以 SNP 形式出现，其中被鉴定出来的已经达到数百万种之多（International HapMap Consortium 2005）。最近的研究也开始强调 CN 多态性频率（Iafrate et al. 2004；Sebat et al. 2004；Redon et al. 2006），这表明全面研究检测样本之间的遗传差异应结合双方的序列和 CN 变异两个方面。

Affymetrix 基因芯片（Affymetrix GeneChip ®）作图阵列包括一类 SNP 寡核苷酸微阵列，可用于确定基因型和在具体 SNP 位置上的 CN。原始作图阵列（10K 作图阵列，Mapping 10K array）以及后来的 10K 2.0 作图阵列在一个单一阵列中查询了超过 10K 的 SNP。下一代绘作阵列扩充到了 100K 作图阵列套装，包括两个阵列（50K Xba 作图阵列和 the Mapping 50K Hind 阵列），可以共同分析 116 204 个 SNP。这些阵列的第三代扩充到 500K 套装，其中包括 250K Nsp 作图阵列和 250K Sty 作图阵列，能检查

SNP 的距离分别为 2.5 kb 和 5.8 kb 的 500 568 个 SNP 中间数和 SNP 间距离。

适用于这些阵列的基因组 DNA 可以用于全基因组抽样分析（WGSA）实验处理（图 16-1）（Kenedy et al. 2003）。重要的是这个实验涉及一个复杂性缩减步骤，因此不是将整个基因组用于这些阵列，而是基因组的一部分被选择性标记和用于杂交。缩减目标 DNA 集合的步骤可以编制如下：在 WGSA 过程中，用限制性内切核酸酶（当使用 500K 作图阵列时选择 NspI 或 StyI）对 250ng 的基因组 DNA 进行消化，并在消化了的 DNA 两端连上接头，在使用这两个 500K 作图阵列时，共需要运用总量 500ng 的 DNA。消化的 DNA 通过 PCR 进行扩增，扩增时采用能识别接头的一套通用引物。PCR 的条件和试剂都进行了优化，以最大限度地扩大特定范围内的扩增产量。对于 10K 和 500K 的作图阵列，这个范围是 100～1100bp；对于 100K 的作图阵列，其扩增子的范围为 250～2kb。

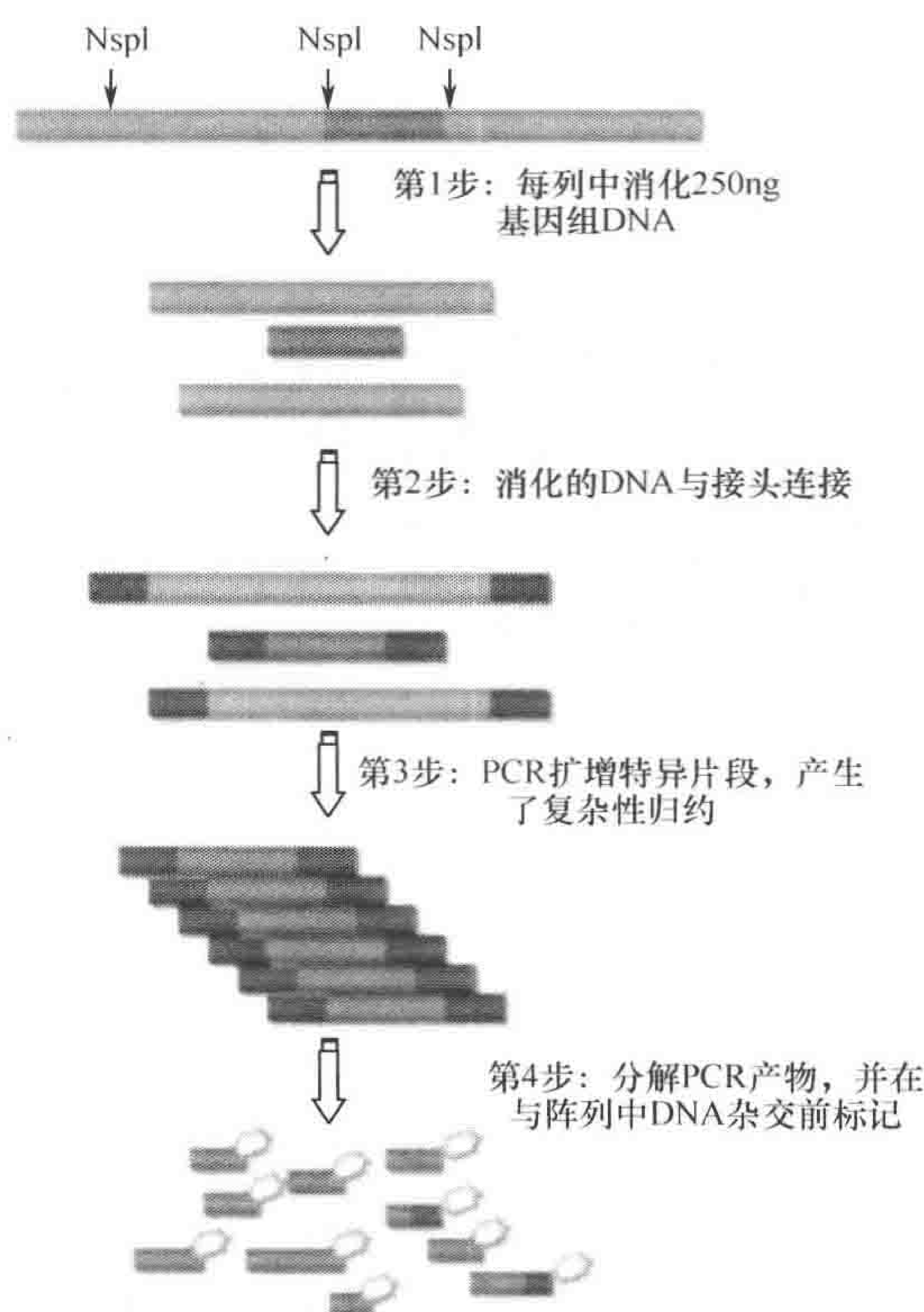



图 16-1　通过全基因组取样分析（WGSA）阵列准备 500K 作图阵列的 DNA

然后，PCR 产物经 DNase I 切割成片段，用生物素标记，并与阵列进行杂交。阵列上识别每个 SNP 的不同等位基因的探针杂交效率通过用荧光标记的 Streptavidin 对生物素标记的 DNA 片段进行染色来检测。

研究者以不同强度的探针来识别主要的和次要的等位基因，这样来达到确定 SNP 基因型的目的（Matsuzaki et al. 2004；Diet al. 2005）。这些探针强度还可以用来测定每一个 SNP 位置上的基因组 CN（Bignell et al. 2004；Huang 2004；Lin

et al. 2004；Nannya et al. 2005）。要确定 CN，需要将测试样本与包含一个或多个样本的对照组进行比较。在此，假定对照组代表了 CN＝2 的一个二倍体状况，计算出试验组和对照组之间的探针强度比，用来确定试验样本中 DNA 片段的拷贝数。

### 用于 CN 检测的 FFPE 样本

WGSA 检测与作图阵列结合，被广泛应用于高质量的 DNA 样本的 CN 分析，即采集自血液、新鲜或冷冻组织或细胞系的样本。不过，福尔马林固定、石蜡包埋（FF-PF）的组织代表了库存临床样本的最通用形式（图 16-2），但同时它给分子水平的实验带来了更多的挑战。FFPE 处理主要有利于病理学家，他们需要快速的手段来固定、染色、查看样本，以进行迅速和高效的诊断。不幸的是，该方法通常造成 FFPE DNA 的降解、污染和化学修饰。

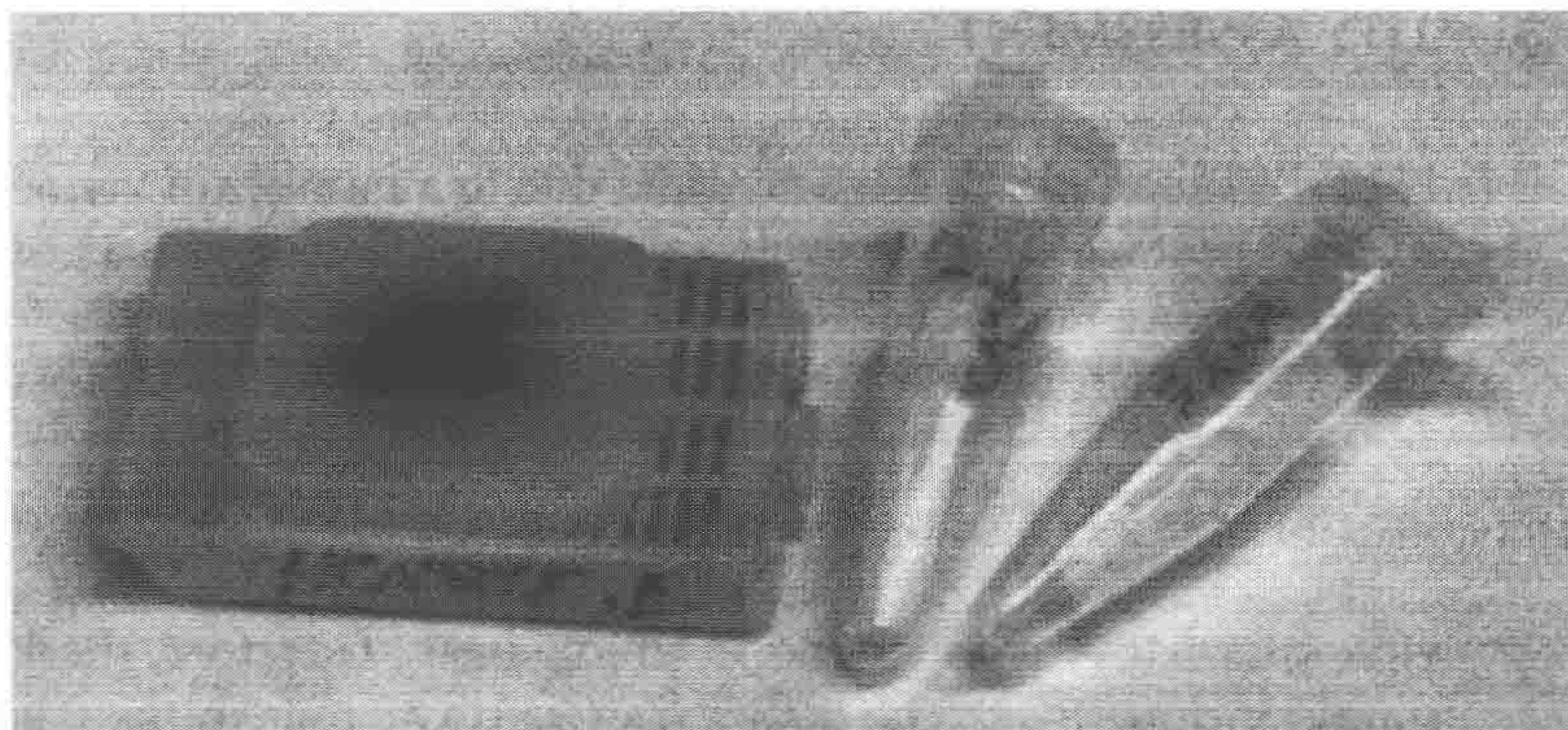


图 16-2　图示含有福尔马林固定的肾脏样本的石蜡块和装有 FFPE 样本的 10μm 切片的 eppendorf 管

由于这些问题，FFPE DNA 并不是适用于所有高品质 DNA 样本的分子检测。下面，我们推荐一项通过 WGSA 处理 FFPE DNA 样本和进行作图阵列分析的程序。请注意，质量较差的 FFPE 样品仍然可能导致 SNP 提供详实数据的数量减少，因此造成整个基因组范围的数据出现问题。FFPE 样品数据分析的分析指南将在第 22 章加以概括。特异适用于 FFPE 导出样本的 WGSA 程序的最重要修改要点包括

（1）包括以前 WGSA PCR 技术为基础的 FFPE DNA 样本质量评估。

（2）收集扩增反应（>3 个）产物，使 PCR 产物达到 90μg 的量值。

（3）质量控制（QC）措施，以监督分析过程运行情况（希望从 FFPE 和非 FFPE 的 DNA 样本得到不同的结果）。

## DNA 样本的制备

目前市场上一些高品质的商品试剂盒可以用来从 FFPE 样本中提取 DNA。从某一

样本提取 DNA 的质量好坏取决于提取程序，因此使用不同试剂盒提取同一 FFPE 样本中的 DNA 可能产生相对较差的结果，也可能获得相对完整的 DNA 样本。一项推荐的经过修正（Wu et al. 2002）的提取过程是用 Qiagen DNeasy 血液和组织试剂盒（www.qiagen.com）来完成的。在 Qiagen Dneasy 血液和组织手册中，相应步骤的修改概括如下。

（1）在 ATL 缓冲液中，在蛋白酶 K 消化前以 95℃温度处理 15 min。

（2）在 56℃，用蛋白酶 K（PK）处理 3d，同时每天补充与第一天添加的等量的 PK。

（3）以 $NH_4OAc$ ＋EtOH 清洗提取的 DNA，方法见《GeneChip® 500K 作图分析手册》中"基因组 DNA 的准备"一节（www.affymetrix.com）。

已经证明能够为阵列作图提供足够优质 FFPE DNA 的另一种方法是 Argylla Technoligies PrepParticles（测试版）（http://argylla.com/）。

有许多可用的提取方法也能够成功获得结果。不论何种方法，由此产生的 DNA 样本都应进行质量评估，因为并不是所有的样品都适合于阵列作图。为了通过质量检验，DNA 样本必须是没有污染和降解的。参阅下一节更详细的 FFPE DNA 样本质量控制测试内容。

## FFPE DNA 的质量评价

由于许多可变因素，FFPE DNA 样本的质量会有所不同，如固定方法、提取方法和存储时间的不同等。一些 FFPE 样品无法产生足够高质量和大数量的核酸用于下游分析。因此，每个 FFPE DNA 样本应进行质量测试，QC 测试不合格的样品就不应该用于阵列作图。不能通过质量测试的部分样本，可能源于 FFPE 样本来源、提取方法以及其他因素不同，其比率也因此发生变化。

在 WGSA 决定一个特定 FFPE DNA 样本是否能够为阵列作图提供详实数据的因素中，PCR 是限制步骤。因此，在 FFPE DNA 质量评估测试中，应包括 PCR。下面表格中的三个以 PCR 为基础的质量控制测试方案，能够确定 FFPE 样本对于阵列作图的适宜度。这三项测试中，任何一个都可用于 FFPE DNA 的质量评价。

| 方法 | 描述 | 优缺点 | 参考文献 |
|---|---|---|---|
| 多重 PCR | 同时扩增若干个大小为 100～800bp 的片段；<br>每个实验室都可以设计引物和反应，用来扩增基因组的任何区域范围；<br>重要的是设计一些反应，既能够扩增小片段又能扩增大片段（600、700bp） | （＋）需要的 DNA 量较小；<br>（－）特异位点靶向的 PCR 扩增反应可能鉴定出的是基因座位特异性的样本质量而非全基因组范围的 DNA 质量 | van Beers 等.（2006），这篇文章仅描述了长达 400bp 多重 PCR |

续表

| 方法 | 描述 | 优缺点 | 参考文献 |
|---|---|---|---|
| RAPD PCR | 以单一一套引物（10mer）来扩增一个梯度的片段；<br>完整的DNA能够产生长达2Kb的扩增子；<br>产生于FFPE样本的扩增子梯度可达到一个较小的最大尺度<br>这不是一个试剂盒 | （+）需要的DNA量较小；<br>（+）提供跨基因组的非定向扩增；<br>（一）如果在制备酶样时未进行高度纯化，Taq聚合酶中的细菌DNA可能被扩增，结果出现较高的背景；<br>（一）在复制产物和稀释样品中可能出现较大的可变性 | Siwoski等（2002），根据建议做了些修正。<br>用EtBr在琼脂糖凝胶中进行条带显色。<br>使用高纯度的Taq聚合酶，如Qiagen HotStar HiFidelity DNA聚合酶 |
| WGSA的头一步反应 | 消化250ng的DNA，连上接头，用一套通用引物进行PCR扩增反应。<br>参见图16-3A作为例子 | （+）由于其本身是作图分析的一部分，提供了与作图分析最好的相关性；<br>（+）通过QC检验的产物可直接继续用于WGSA分析，与阵列进行杂交。<br>（一）每个样品需要最大量的DNA、最长的时间和最多的费用 | GeneChip® 500K作图分析手册 |

这些质量控制测试的目的，是为了确认对某一给定FFPE样品，可产生的最大扩增子的大小。重要的是，这些QC例子都包含了PCR中扩增的各种大小的片段，包括800bp的较大片段大小。因为500K作图阵列能解读大到1100bp的片段的SNP，一个适合阵列分析的FFPE样本应能够提供较大的扩增子大小。虽然没有必要扩增阵列上的所有片段（在芯片上，最大片段上的SNP可以排除在分析之外），重要的是要扩增显著比率的这些SNP。表16-1指出了在不同片段大小上SNP的数目。我们不建议对未能显示大于300 bp扩增片段的样本继续进行实验。

**表16-1　使用不同片段大小滤器（filter）时驻留在500K作图阵列上的SNP的数目**

| 滤器适合的片段大小 | 250K Sty作图阵列 | 500K作图阵列 | 组合500K套装 | 500K套装上的SNP/% |
|---|---|---|---|---|
| ≤200 bp | 1 260 | 1 813 | 3 073 | 0.6 |
| ≤300 bp | 15 831 | 13 650 | 29 481 | 5.9 |
| ≤400 bp | 45 459 | 39 506 | 84 965 | 17.0 |
| ≤500 bp | 82 085 | 74 387 | 156 472 | 31.3 |
| ≤600 bp | 120 011 | 113 702 | 233 713 | 46.7 |
| ≤700 bp | 155 576 | 153 213 | 308 789 | 61.7 |
| ≤800 bp | 187 689 | 190 900 | 378 589 | 75.6 |
| ≤900 bp | 213 302 | 222 317 | 435 619 | 87.0 |
| ≤1 kb | 230 509 | 244 665 | 475 174 | 94.9 |
| No filter | 238 300 | 262 258 | 500 568 | 100.0 |

## 应用 FFPE DNA 于 WGSA 分析中

在 FFPE DNA 处理时，一个 500K 作图套装或者一个单一的 500K 作图阵列，应优先于其他作图阵列而被采用。100K 的作图阵列不适合降解的样本，这是由于它在 PCR 扩增反应时要求的扩增子较大。10K 的作图阵列可用于 FFPE DNA 分析，因为它有着与 500K 作图阵列同样的 PCR 扩增子的大小分布，但是这种阵列将显著地降低分辨率。

对于 WGSA 实验，Affymetrix（www. aggymetrix. com）提供有详细的 GeneChip ® 500K 作图分析手册。在这里，我们强调一些需要特别注意的或在 FFPE DNA 样品应用中进行了修改的具体步骤。

### DNA 定量

为准确地对双链 FFPE DNA 进行定量，需要使用一种特异检测双链 DNA 的方法，如 PicoGree（www. invitrogen. com）。还有其他定量方法，如紫外吸收光谱法，由于 FFPE 样本含较大量的污染物，可能提供不精确的 DNA 测量数据。

### DNA 的消化和连接

在 WGSA 的消化和连接中，不需要做任何操作步骤改变。从 250ng 的 DNA 用量开始，用 NspI 消化样品以达到进行 250K NSP 阵列作图的目的，用 StyI 消化样品用于 250K Sty 阵列作图分析。完成连接反应的结果，将得到 100μL 的产物。

### PCR

已有的实验步骤要求每个 DNA 样本要进行 3 个扩增反应，每个 PCR 反应要用 10μL 连接产物。因此，总共要从 100μL 连接产物中使用 30μL。PCR 扩增后，将这三个反应产物收集起来，将 90μg 的 PCR 产物使用到下一个步骤（片段化）。由于降解的 FFPE DNA 在 PCR 中可能不会产生较大的扩增子（图 16-3A），PCR 的产物收率往往低于 FFPE 样品的情况，因此有必要收集额外的 PCR 产物，以达到所要求的 90μg。由于预计需要更多的 PCR，每个样品设定 6 个扩增反应（从 100μL 连接产物中使用 60μL）。在严重降解的样本中，可能需要 9 个反应。重要的是，这一步骤不会影响 DNA 的输入量，每个阵列仍然是 250ng。

一个 FFPE 样品只组合三个 PCRs 产物是可以接受的，即使这提供不了 90μg 的产物量，但集中更多的样本使更多的 DNA 应用到阵列中，一般能够提高结果数据的质量。

### 纯化和汇集扩增反应

用 Clontech ® 96 孔板来进行 500K 作图阵列的 PCR 的收集和纯化。在一个孔中可以集中多达 3 个反应量。如果对于同一 DNA 样本进行超过 3 个 PCR 反应，应该用多个

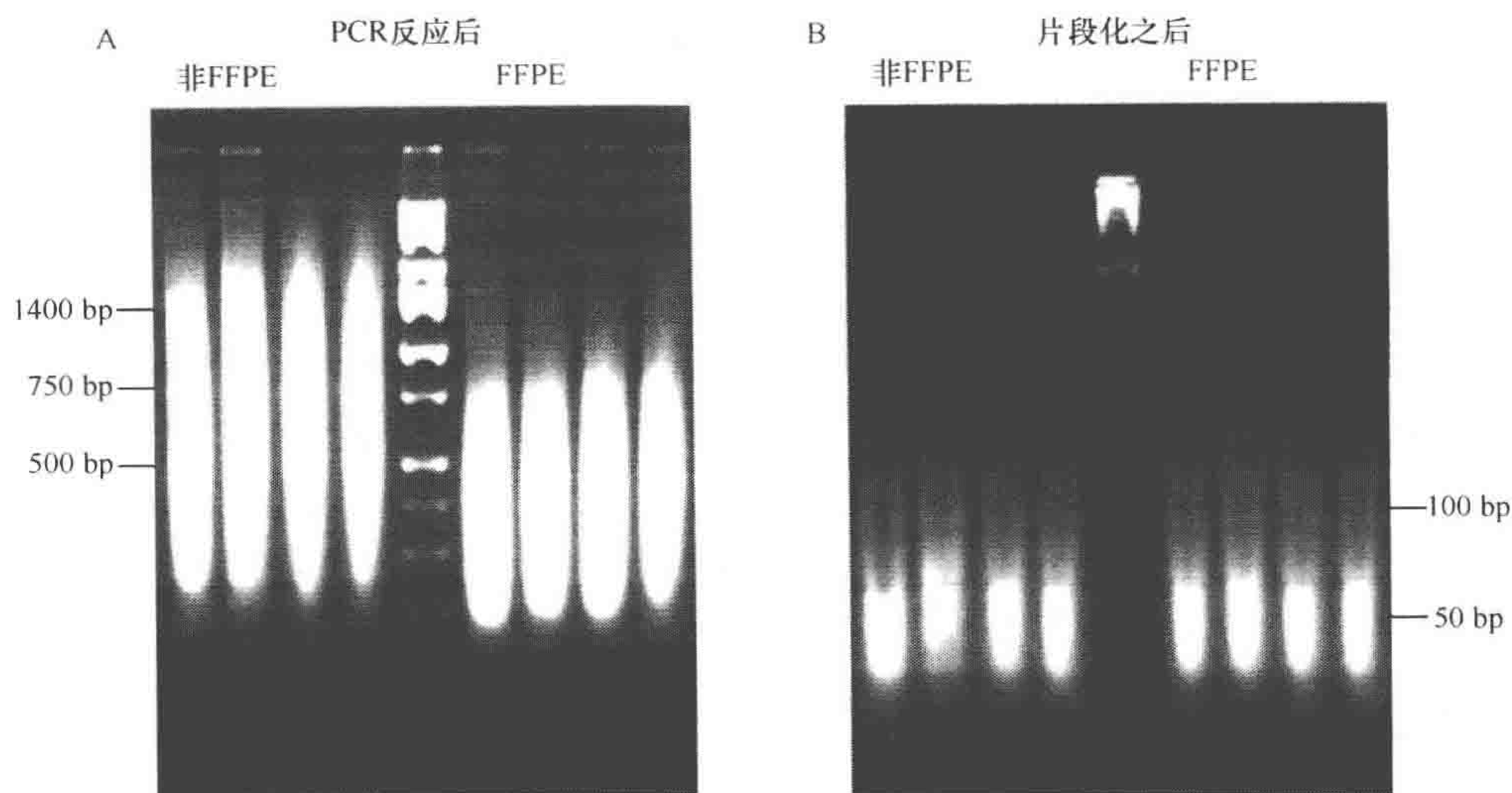



图 16-3　WGSA 反应体系中样本的 QC 检测。A. 在琼脂糖凝胶上对非 FFPE 和 FFPE 样本的 PCR 产物进行显示。在 FFPE 反应中，大片段没有能够被扩增出来；不仅如此，还出现了大量的较小片段 PCR 产物，这说明这些样本适合于用在 WGSA 反应中，并与阵列进行杂交。B. 在琼脂糖凝胶上对非 FFPE 和 FFPE 样本的片段化产物进行显示。产物看来是等量的，适合于标记和用于与阵列杂交

孔来纯化。因此，如果同一来源 DNA 有 6 个 PCR 反应，将用 2 个孔用于这个样本 PCR 反应产物的纯化。从这些孔中得到的 DNA 可以经过以下洗脱步骤后进行合并：在用水清洗每个孔 3 次以后，让 DNA 在平板上完全干燥（按照 GeneChip ® 500K 作图分析手册中的指示），用 45μL RB 缓冲液洗脱第一个孔，而留下第二孔保持干燥。然后，使用第一个孔中含有 DNA 的缓冲液，对第二个孔进行洗涤。通过这种方式，所有 6 个反应的 PCR 产物将洗脱到同样的 45μLRB 缓冲液中。如果同一来源 DNA 有 9 个 PCR 反应，对于这个样品的纯化要使用 Clontech ® 孔板中的 3 个孔，以此类推。

### 片段化和标记

可能的话，对 90μg 的 PCR 产物进行片段化。如果较少量的 DNA 作为片段化的输入样品，可能使阵列上反应减少。

### 数据分析

参阅第 22 章中关于如何分析从 FFPE 样品取得阵列作图数据的细节。

## 性能指标

DNA 样本的性能可在 WGSA 实验过程中或者之后进行监测（见下表）。在一批 FFPE 样本中包括一个好的对照是很有用的，如由 Affymetrix 试剂盒 Ref103 提供的高

品质标准 DNA 样本。

| QC 检测 | 细节描述 |
|---|---|
| 在琼脂糖凝胶上显示 PCR 产物图谱（图 16-3A）<br>用紫外分光光度仪测量 PCR 产量 | 非 FFPE 样本：PCR 产物应该显示为从接近 100bp 到>1kb 的弥散分布图谱。<br>FFPE 样本：较大的 PCR 产物可能丢失，但较小的 PCR 产物应该获得充分的扩增。不要采用未显示大于 300bp 的强带的样品<br>非 FFPE 样本：在一个 3 孔反应体系的收集量中，应该能提供>90μg 的产量。<br>FFPE 样本：这样的样本可能显著减少 PCR 产物的产出。理想的情况下，当 6 个或 9 个 PCR 体系汇集到一起时，样本能提供的产物量应该在 90μg 规模。如果 9 个反应未能获得 90μg PCR 产物，就用这些样品做最小限度的反应。一般来说，较大的产量意味着在数据分析中能够包括的较大数量的 SNP |
| 在琼脂糖凝胶上显示片段化产物图谱（图 16-3B） | 片段化产物应该为 50～100bp，不管是对于非 FFPE 样本还是 FFPE 样本 |
| 检查收益率（call rate） | 非 FFPE 样本：在利用 DM 算法时，当收益率（SNP 分布的基因型的比率）≥93%* 时可以得到最好的数据；而在利用 BRLMM 算法时，是≥96%**。两种算法都在基因型分析软件 GTYPE4.1 中使用。<br>FFPE 样本：这样的样本可能显著减少收益率。而且，要成功应用时，收益率需要达到 70%以上。如果更低，数据质量就会很差 |

*0.33 的 $p$ 值 cutoff。

**0.33 的 $p$ 值 cutoff 和 0.5 的 BRLMM 得分阈值。

# 结论

对于存档的临床标本，FFPE 样品代表了最大量的资源；但由于 DNA 降解、化学修饰，以及样品污染，它们也给分子实验带来了特殊的挑战。想要从给定的一个 FFPE 样品中尽可能提取高品质的 DNA，提取方法应给予特殊考虑。为了确保优质，所有 FFPE 样品应采用以 PCR 技术为基础的测试。那些提供足够扩增量的样品，可通过 500K 作图阵列用于全基因组基因型和 CN 分析，覆盖全基因组的 FFPE 样品与高品质的 DNA 样本进行比较，可以减少 PCR 技术在大片段扩增 SNP 时失败的情况。

对于应用 FFPE 样本进行作图分析，很少有修正方案可以推荐。在 WGSA 过程中，在片段化反应之前增加扩增反应的数量以达到 90μg 的 PCR 产物，一般都能提升数据质量。在数据分析过程中进行一定的方案修改也是必要的，我们将在本手册的第 22 章加以讨论。当使用高质量的 DNA 样品时，珍贵的样品的全基因组扩增可以与 WGSA 试验配套进行（Wong et al. 2004；Zhou et al. 2005），但是在此时，如果在用于作图分析之前样品已经降解，那就没有可借鉴的经验让你成功地完成全基因组扩增了。

# 参考文献


Bignell G.R., Huang J., Greshock J., Watt S., Butler A., West S., Grigorova M., Jones K.W., Wei W., Stratton M.R., et al. 2004. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14:** 287–295.

Di X., Matsuzaki H., Webster T.A., Hubbell E., Liu G., Dong S., Bartell D., Huang J., Chiles R., Yang G., et al. 2005. Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* **21:** 1958–1963.

Huang J. 2004. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics* **1:** 287–299.

Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., Scherer S.W., and Lee C. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* **36:** 949–951.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Kennedy G.C., Matsuzaki H., Dong S., Liu W.M., Huang J., Liu G., Su X., Cao M., Chen W., Zhang J., et al. 2003. Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21:** 1233–1237.

Lin M., Wei L.J., Sellers W.R., Lieberfarb M., Wong W.H., and Li C. 2004. dChipSNP: Significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20:** 1233–1240.

Matsuzaki H., Dong S., Loi H., Di X., Liu G., Hubbell, E., Law J., Berntsen T., Chadha M., Hui H., et al. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1:** 109–111.

Nannya Y., Sanada M., Nakazaki K., Hosoya N., Wang L., Hangaishi A., Kurokawa M., Chiba S., Bailey D.K., Kennedy G.C., and Ogawa S. 2005. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.* **65:** 6071–6079.

Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shapero M.H., Carson A.R., Chen W., et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.

Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Maner S., Massa H., Walker M., Chi M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Siwoski A., Ishkanian A., Garnis C., Zhang L., Rosin M., and Lam W.L. 2002. An efficient method for the assessment of DNA quality of archival microdissected specimens. *Mod. Pathol.* **15:** 889–892.

van Beers E.H., Joosse S.A., Ligtenberg M.J., Fles R., Hogervorst F.B., Verhoef S., and Nederlof P.M. 2006. A multiplex PCR predictor for aCGH success of FFPE samples. *Br. J. Cancer* **94:** 333–337.

Wong K.K., Tsang Y.T., Shen J., Cheng R.S., Chang Y.M., Man T.K., and Lau C.C. 2004. Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. *Nucleic Acids Res.* **32:** e69.

Wu L., Patten N., Yamashiro C.T., and Chui B. 2002. Extraction and amplification of DNA from formalin-fixed, paraffin-embedded tissues. *Appl. Immunohistochem. Mol. Morphol.* **10:** 269–274.

Zhou X., Temam S., Chen Z., Ye H., Mao L., and Wong D.T. 2005. Allelic imbalance analysis of oral tongue squamous cell carcinoma by high-density single nucleotide polymorphism arrays using whole-genome amplified DNA. *Hum. Genet.* **118:** 504–507.

# 17 分子倒位探针靶向的基因型分析——拷贝数确定的应用

George Karlin-Neumann, Marina Sedova, Ronald Sapolsky, Steven Lin, Yuker Wang, Martin Moorhead, and Malek Faham

*Affymetrix, South San Francisco, California 94080*



## 简介

发现和精确地测量癌症样本中的染色体拷贝数目（CN）扩增、缺失以及杂合性丢失（LOH）的能力，对于诊断来说具有很大的潜在临床意义（Slamon et al. 2001）。在染色体拷贝数目扩增的确定中面临的一个基本的挑战是在基因型测量中不仅要有能力达到定性的目的，还要有能力获得定量指标。对于癌症样本来说，还有样品纯化难（因为待测组织中癌细胞与正常细胞鱼龙混杂）、DNA 样品含量低，以及有时保存用的福尔马林固定、石蜡包埋（FFPE）样本、物理或化学降解的 DNA 带来的麻烦，使得扩增效果不佳、用现有的手段难以分析（van Beers et al. 2006）。人们开发了几种以阵列为基础的技术，用以估计 DNA 的拷贝数目。理想的拷贝数分析能够提供多样化的基因型分析风格、精确的拷贝数估计，不光是适用于冗余的、高质量的 DNA 样本，而且也可用于含量低、有降解的 DNA。第 12 章描述了在目标基因型分析中分子倒位探针（MIP）的用途和优点。在拷贝数研究中运用 MIP 的优点源自其良好的特异性，因为需要两个处于同样分子上的同源序列发生相互作用，如第 12 章的描述那样。另外，由于需要的靶序列小（带探针末端检测序列的～40 bp 的同源性），它能够适于对有一定降解的样品进行分析。这些特征使得 MIP-基因型分析技术能够很好地适应拷贝数确定方面的挑战，即使对于低含量、未扩增和有降解的样品 DNA。MIP 分析的定量性能在下文会描述，其中会对实验操作进行一些修正，不是来自商售的。这一分析版本需要＜75 ng 的未扩增的、并且已经用一个 50 000-plex MIP 基因组范围的人类样品组对来自正常和癌细胞系样本以及 FFPE 样本的 DNA 进行测试的基因组 DNA。它对每个样本的分析性能都很好。为了确定一份已知样本中给定标记的每种等位基因的拷贝数，等位基因各自的信号强度首先是针对一套对照样本标定好的。

## 来自 MIP 基因型检测的 CN 结果评估

为了评价 MIP CN 评估的准确度和精确度，用不同拷贝份数的 X 染色体细胞系进行试验，从男性的 1X 到 3X、4X 和 5X 细胞系（细胞系来自新泽西 Camden 的 Coriell 研究所）。在 X 染色体上 50K-plex panel（50K 丛面板）可包含多于 900 个可评价标记。标记的 CN 与染色体的对比绘制成图见图 17-1，示 1X、3X、4X 和 5X 细胞系的情况。每条染色体中标记清楚地被染色，从而与相邻位置进行区分，常染色体从左至右排列，从 1 号染色体（左面，红色）到过 22 号染色体（右面，粉红色），同时 X 染色体的标记在最右端（桃红）。每个标记 CN 值并不光滑，但尽管如此，却显示出很高的精确性；仅有的两个平滑的标记进一步增加了这一精确度。在 1X 的男性中 X 染色体标记的 CN 评估值为 1.06，相对标准偏差（rsd）为 0.12；3X 的细胞系，CN=3.09，rsd 为 0.10；4X 细胞系，CN = 3.98，rsd 为 0.1；5X 细胞系，CN = 4.96，rsd 为 0.10。拷贝数的评估在此范围内是准确和精确的。值得注意的是，这个实验在这些细胞株中已经能够识别 CN 异常的一些推测地区，如在 3X 的细胞系中 14 号染色体（紫色）出现明显扩增。更多具有已知浓度峰值的数据表明，分析结果在含 50 份以上拷贝时是线性的（数据未显示）。

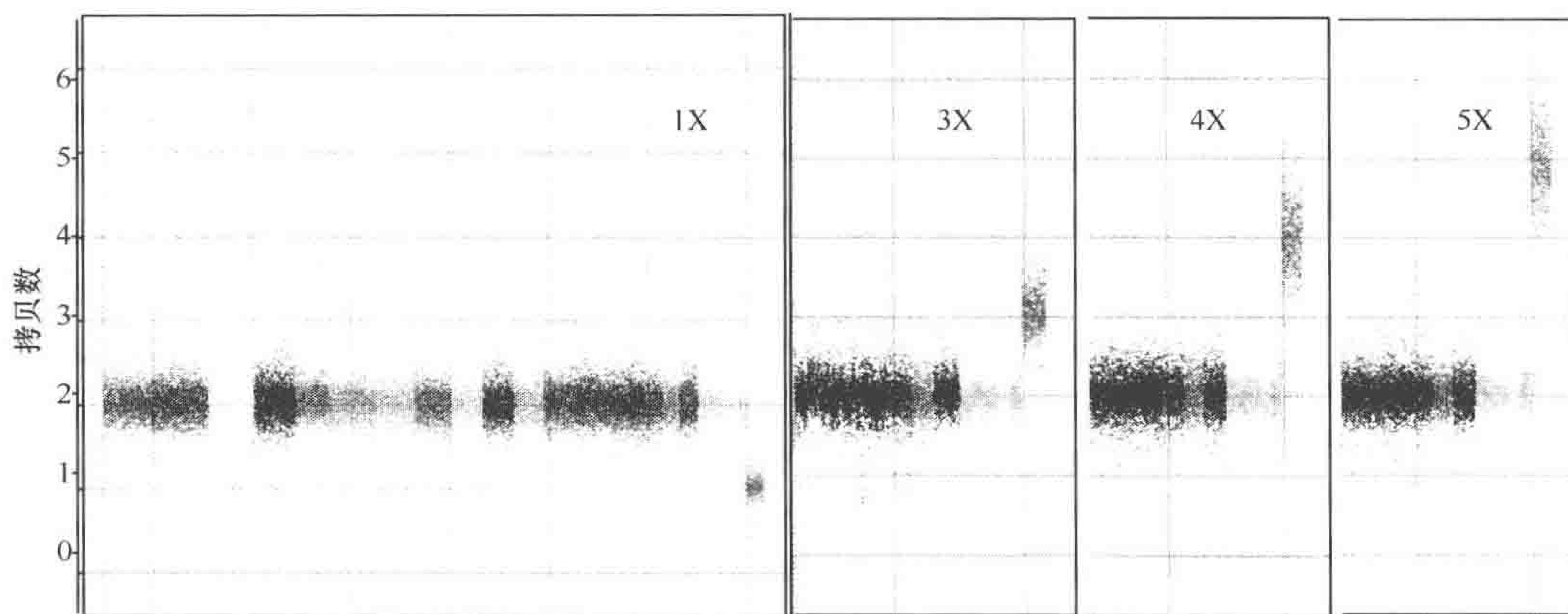



图 17-1 在 X 染色体滴定细胞系中，CN 度量的精确度和准确度。4 个图形从左到右依次对应于 1X（雄性）、3X、4X 和 5X 细胞系。在每一个图中，常染色体从左至右排列，从 1 号染色体（左面，红色）到 22 号染色体（右面，粉红色），同时 X 染色体的标记在最右端（桃红）。代表了两种等位基因的 CN 总和，来自非光滑的 CN 度量。详细讨论见正文（见图版）

这个检测在确定此范围内的缺失区域和扩增区域方面的能力如图 17-2 的进一步描述。此图描绘了癌细胞株 UACC812［细胞系从美国菌种保藏中心（ATCC）获得］的染色体 CN 图。染色体从左至右排列，1 号染色体在最左边，X 染色体在右边远处。几个标记的拷贝数惊人的增加是显而易见的。其中有几项扩增的水平也已通过 TaqMan 测量证实（Wang et al. 2005）。

在这一应用的第三个例子中，我们研究了这一修改的 MIP 实验很好地运用在来自

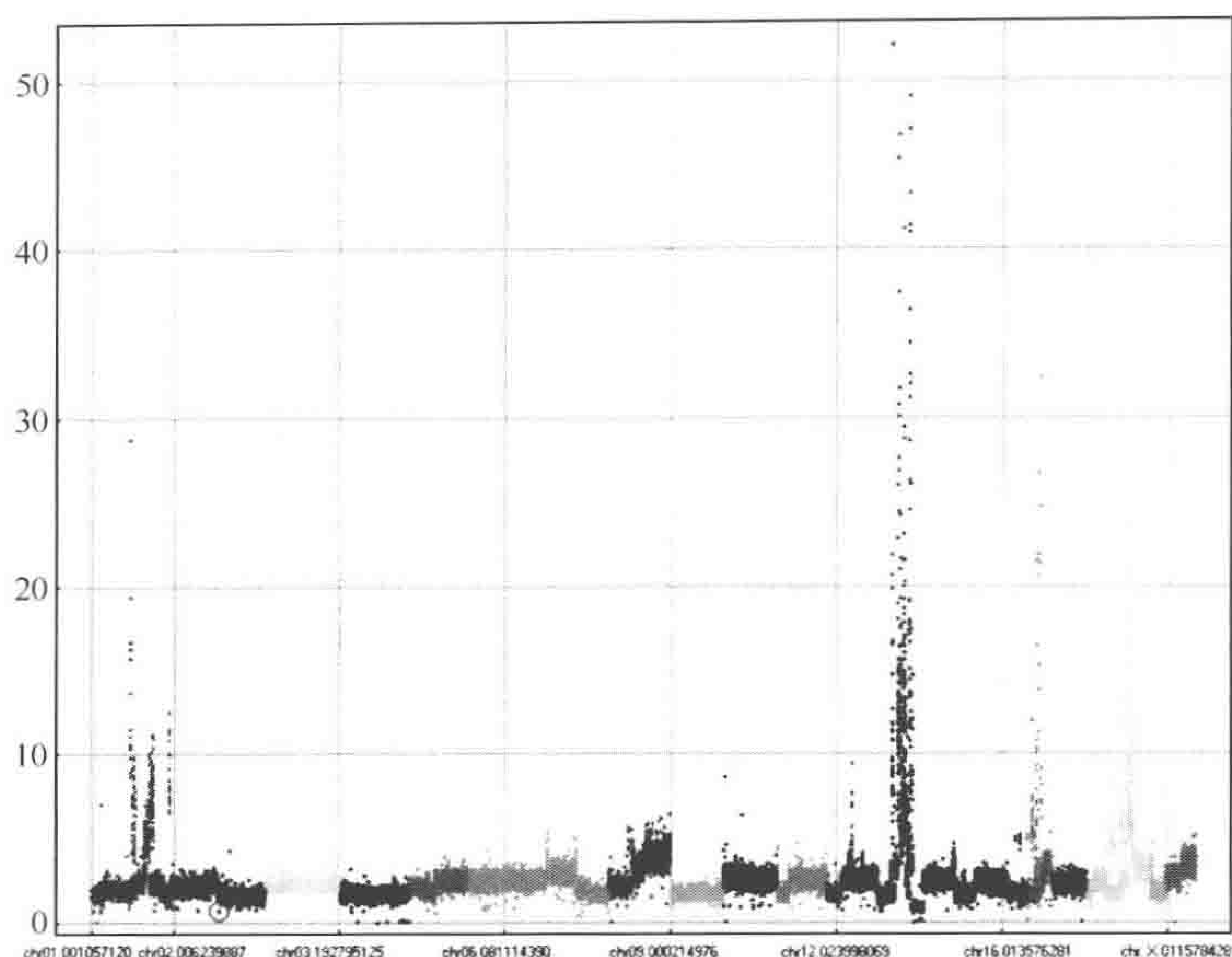


图 17-2 UACC812 癌细胞系的扩增和缺失。在此图中，这一细胞系中染色体也是自左向右排列，一系列显色标记依次表示 1～22 号染色体。着色的画图轮廓描述了相应标记的扩增或缺失。所有的点表示两个等位基因的 CN 加和，来自光滑的 CN 量度

正常和肿瘤的 FFPE 样本 DNA 上的能力。在这种情况下，DNA 样本利用标准 QIAmp DNA 提纯程序（Qiagen）纯化所有样品块的若干 10μm 切片，并在上述对于细胞系 DNA 描述相同的条件下进行分析。从一个具有代表性的正常男性 FFPE 样品中获得的 CN 对染色体的作图结果在图 17-3 中展示。同图 17-1 显示的细胞系区一样，每一点代表了一个单一非光滑标记的 CN 测量值。在图 17-3 中，作图区域不仅显示已知的单拷贝 X 染色体标记的准确定量（桃红色，右侧端部区域），还很好地与双拷贝常染色体标记进行区别（保持代表 1～22 号染色体一系列彩色标记）。因此，MIP 基因型检测实验可以为正常的和 FFPE 样本来源的 DNA 提供高品质的 CN 测量。

## 概要和结论

除了提供高度准确的基因型外，MIP 检测系统的基本性能是能够实现成套的准确和精确的高度重复标记拷贝数测量。由于此检测的高特异性和广泛的动态范围，缺失和扩增的准确测量几乎可以达到近 2 个数量级。由于等位基因之间没有干扰，可以进行等位基因特异性拷贝数的精确测定而不会有相邻标记的滤波。MIP 探针在其基因组靶位（～40bp）上的小印记使其非常适合测量来自 FFPE 样中的降解 DNA 拷贝数。这在 50 000丛（plex）水平上已被证明，从而使这项技术适用于进行全基因组的缺失、扩增和 LOH 评估。

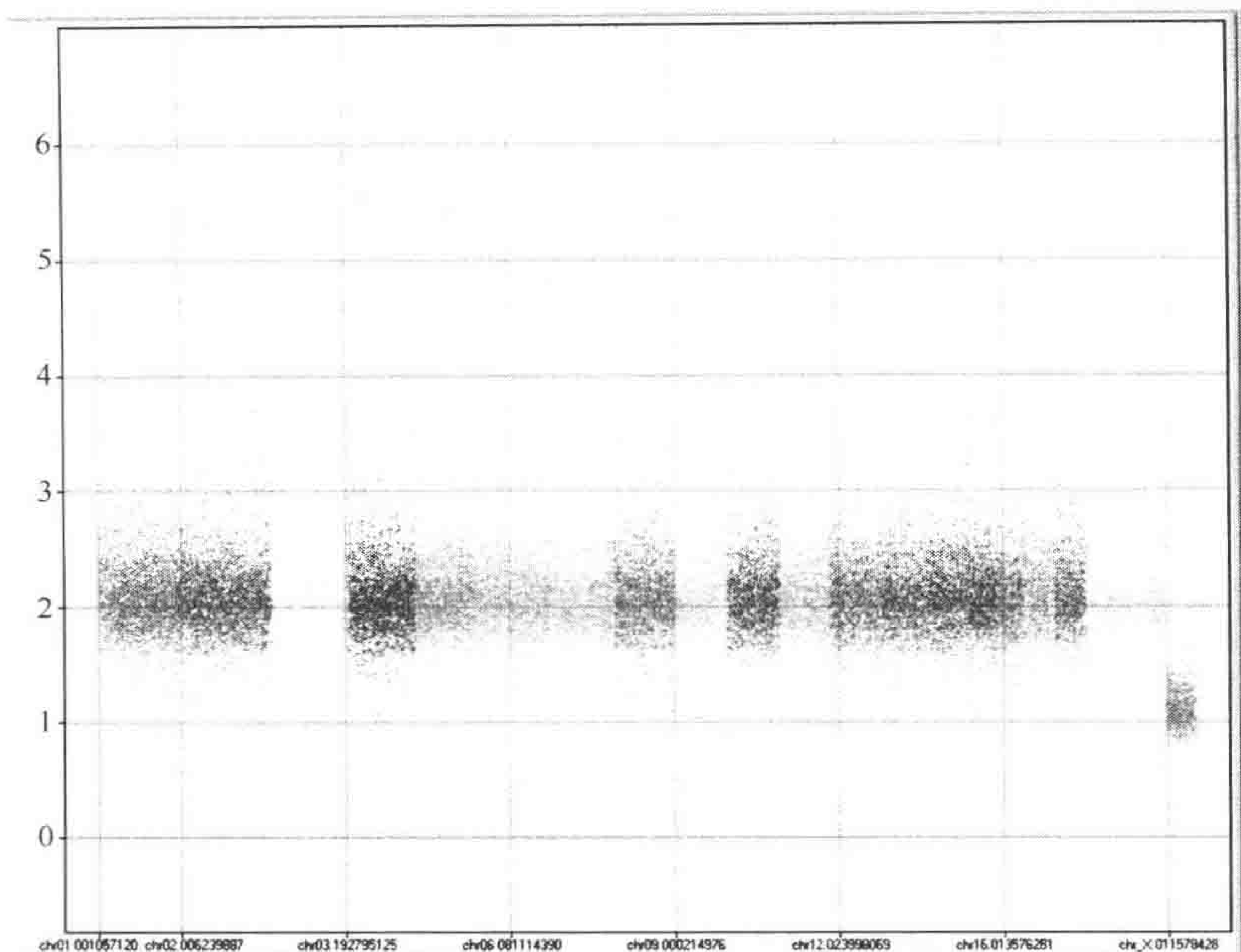



图 17-3　正常男性 FFPE 组织（1X 对 2X）的 CN 作图。DNA 样本利用标准 QIAmp DNA 提纯程序（Qiagen）纯化自样本 FFPE 组织块的若干 10-μm 切片。染色体的组织见图 17.1 的描述（2X 常染色体自左到右排列，1X 染色体在远侧右端），CN 基因型确定分析在图 17-1 对细胞系描述的相同条件下进行。所有的点代表两个等位基因的 CN 加和，来自非光滑的 CN 量度。结果的细节在下述正文中讨论（见图版）

## 致谢

我们感谢来自 ParAllele 和 Affymetrix 的一直致力于发展 MIP 技术的许多人士，特别是通过运算和软件开发帮助拷贝数分析的人。

## 参考文献

Slamon D.J., Leyland-Jones B., Shak S., Fuchs H., Paton V., Bajamonde A., Fleming T., Eiermann W., Wolter J., Pegram M., et al. 2001. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344:** 783–792.

van Beers E.H., Joosse S.A., Ligtenberg M.J., Fles R., Hogervorst F.B., Verhoef S., and Nederlof P.M. 2006. A multiplex PCR predictor for aCGH success of FFPE samples. *Br. J. Cancer* **94:** 333–337.

Wang Y., Moorhead M., Karlin-Neumann G., Falkowski M., Chen C., Siddiqui F., Davis R.W., Willis T.D., and Faham M. 2005. Allele quantification using molecular inversion probes (MIP). *Nucleic Acid Res.* **33:** e183.

# 18 微卫星标记的连锁和关联研究

Jeffrey Gulcher

*deCODE Genetics, Reykjavik, Iceland, and Woodridge, Illinois*



## 简介

20 世纪 90 年代和 21 世纪开始的几年，在独立于假说的人类基因组研究中，微卫星或短串联重复序列（STR）成为极受关注的重要遗传标记，促进了全基因组连锁和等位基因不平衡研究。然而，随着高通量和符合成本效益的 SNP 平台的崛起，我们目前正处于用 SNP 进行基因组扫描的时代。然而，必须指出的是，对于连锁和相关研究来说，微卫星仍然是具有较大信息量的、很有用的基因变异检测措施。在种群历史上，它们补充 SNP 的持续优势在于它们有更大的等位基因多态性而非二等位基因 SNP，其中串联重复序列的单步扩张或收缩在祖先 SNP 的单体型背景下，可以打破共同单体型，从而在感兴趣的连锁不平衡区（LD block，LD 功能块）导致更大的单体型多样性。事实上，微卫星最近在相关研究中扮演主角，使得人们在大范围内接连地发现 2 型糖尿病（TCF7L2）和前列腺癌的基因（the 8q211 region）（Amundadottir et al. 2006；Grant et al. 2006）。最后，对所有变异的目录进行造册也是很重要的，包括 SNP、微卫星、拷贝数变异，以及在人类遗传研究中的倒置多态性。本章描述的是微卫星的作用以及它们在实验方法中的应用。

### 在连锁研究中微卫星的应用

由于种种原因，微卫星成为连锁研究中选择的变量。第一，它具有等位基因多态性，平均 5～10 个等位基因有一个变异。这就导致了普遍应用的微卫星杂合率达到 50%～80%；与此相反，二等位基因标记（如 SNP）的杂合率最可能为 50%，但实践中，SNP 的平均杂合度为 30%或更低。在杂合率上这样的增加使得相同数量的微卫星比 SNP 能提供更多的信息内容。对于连锁，SNP 用于连锁分析的提倡者的经验表明，

如果所做的是多点连锁分析，应用的 SNP 数量越大（比微卫星多 5～10 倍），SNP 为连锁分析提供的信息就越翔实。如果 SNP 可以被可靠地定型与质量控制、在密度上不是太高以至于导致 SNP 标记间的 LD 的话，这当然是真的。如果不能确定连锁标记不在给人深刻印象的有人为迹象的连锁峰的 LD 结果中，那么连锁程序能呈现出所有标记是处于连锁平衡的。

第二，微卫星能够自一个小量 DNA 用尝试扩增的方法由聚丙烯酰胺分离高通量放射性片段，它是在这个 PCR 系统中进行测量的第一变量，代替了用至 20 世纪 90 年代早期的比较笨重的依赖于 Southern blot 的 RFLP。在通量量级上的进一步扩大，是代替放射性标记的 PCR 片段的荧光标记，并采用片段的自动化实时阅读方式来判读通过 Applied Biosystem 系列分析仪分离的片段（Davies et al. 1994）。荧光标记技术使得多重标记信号能够以不同的颜色得以标记，并且在电泳等展示图谱的相同泳道上以不同的位置显示出长度差异，同时显示了板状凝胶上每个泳道的内部标准相对分子质量信息，大大地增加了片段长度评估的准确度。DNA 与 PCR 试剂的智能化混合和等位基因记录软件系统的开发利用，已经适当地跟上了这种提升了通量的技术发展（Palsson et al. 1991）。在 20 世纪 90 年代末以前，这种高通量技术因毛细管电泳技术的问世而迅猛发展，这使得上样技术实现自动化，并且避免了不同泳道的交叉污染和交叉漂白现象，而这些麻烦在凝胶电泳实验中有时很难克服。一个 ABI 3730 序列分析仪目前每天至少可以产生 30 000 个基因型的分析结果。这对于现行的 SNP 芯片技术来说显得相当合适。但是，我们还应该清醒地认识到，如果要考虑每个标记所增加的信息容量，那么相对于基于定做样本的 SNP 定型分析来说，使用微卫星技术在费用上仍面临很大挑战。

第三，在质量控制方面，微卫星技术很容易做到。就像在 SNP 技术中那样，微卫星技术时而可能因 SNP 处于引物位点内而遭受等位基因中途丢失的问题（也就是说，有时会检测不到两个等位基因中的某一个，使得在纯合子基因型的判定上出现错误）。在微卫星和 SNP 之间，这样的现象所发生的比率大体相当。不过，等位基因中途丢失问题发生在 SNP 中，却要比发生在微卫星当中难以检测到，结果是在杂合子评判上出现很大差异。在相同数量的 SNP 和微卫星当中，前者因等位基因中途丢失而发生基因型判断失误的概率相当于后者的 3 倍之多。如果对于一组已知的谱系结构，要求用于产生良好的信息量（90%）的 SNP 数量是产生相同信息量水平的微卫星数量的 5～10 倍，那么，遭受检测不到基因型的误差的 SNP 标记数目将可能是在等量微卫星扫描中遭受这样误差的微卫星标记数量的 15～30 倍。

第四，在微卫星分析中，能产生高分辨率的遗传图谱，而在 SNP 分析中得不到。2002 年发表的 deCODE Genetics High-Resolution Genetic Map（HRGM）（《deCODE Genetics 高分辨率遗传图谱》）（Kong et al. 2002），其工作是运用 1257 个减数分裂事件［平均分辨率为 0.6 分摩尔根（cM）］在 5136 个经专业筛滤的微卫星标记规模中完成的。还有一种 Marshfield 图谱，其产生运用了超过 8000 个微卫星的标记数量，但分辨率很低（3 cM），原因是谱系中的减数分裂事件的数量只有 188 个。Illumina 技术产生的图谱运用的是 4000 个 SNP 标记，但是所使用的家族的数量是具有与 Marshfield 图谱相似数量减数分裂事件的规模（Murray et al. 2004）。有的分析还做了尝试，通过基

于 deCODE 微卫星标记的物理位置涵盖 SNP 的方式将遗传距离分析信息填写进去。然而，由于物理距离几乎完全不与遗传距离成正比，在遗传距离的评估上可能产生很显著的误差，这可能导致连锁信号的正负两个方面的分析错误（Gretarsdottir et al. 2002）。

微卫星技术并不是一种简单便捷的基因型分析技术，能够有效确立平台和合理利用微卫星分析手段的实验室并不太多。幸运的是，有几家能进行高通量分析的实验室能够熟练地进行性价比合理的微卫星基因型分析，并很少失误（错误少于 0.3%～1%）。Marshfield 临床实验室是一家由微卫星基因型分析技术的主要开发人 James Weber 领导的实验室，在高通量微卫星基因型分析服务行业是开路先锋（Weber and Broman 2001）。Weber 最近已经转聘于一家私营公司（www. preventiongenetics. com)。CIDR 是另一个基因型分析服务资源，虽说它集中处理 10 cM 框架的标记的扫描分析并且不定做将起始连锁信号穷追到底的微卫星样本（www. cidr. jhmi. edu)。最大的、能提供完整服务的微卫星基因型分析实验室要数作者的 deCODE Genetics 实验室了，可以应用 500、1000、2000 和 5000 个微卫星标记提供高密度的基因组扫描分析（www. decode. com/genotyping)。它也提供区域特异性的高密度面板，用来在起始连锁峰内选择性地提高信息量，同时能够发掘有效的微卫星标记定做材料、进行关联研究相关检测（见下文）。

综上所述，微卫星已经成为在连锁研究中最广泛地应用的变异，甚至在如今高密度 SNP 基因型分析时代仍显优势。受本章内容所限，它在连锁分析中的应用方法不再赘述，但在 Gulcher 和 Stefanson（2006）的论文中可以阅读到。

## 关联（association）研究中的微卫星运用

现已清楚，基因中大多数引起普通疾病的功能性变异并不是发现在编码的外显子中的非对称 SNP。相反，它们好像对基因表达和 RNA 剪接有一定影响。来自 HapMap 计划的宝贵数据，使得非冗余 SNP 亚群的选择进行得更加谨慎，使得每个 LD 功能块的标签非依赖性单体型（tag-SNP）作为功能性变异的替代物。到目前为止，利用 HapMap 数据的基因组范围的高密度 SNP 芯片的唯一卖主是 Illumina，利用它可以确定和选择适合于其 300 K、550 K、650 K 和 1 M SNP 珠子芯片的 tag-SNP。这与随机 SNP 选择的不同可甄选方法形成了对比，正如供应者（如 Affymetrix）所实践的那样。

在第 19 章，我们将详细叙述 tag-SNP 选择内容。利用 tag-SNP 进行全基因组扫描的最终结果可以达到 LD 功能块的确定，其中 SNP 高度关联，而单体型多样性相当低。采用微卫星技术，就能够克服限制单体型在 LD 功能块中空间单体型有限多样性这个潜在不利条件。我们注意到许多研究组报道了在每个 LD 功能块中运用几个 tag-SNP 的方式，并发现这些 SNP 限制了 LD 功能块中至少九成的单体型。然而，认识到单体型的涵盖范围不等于单体型多样性非常重要。因此，如果标志性等位基因的频率远远高于它作为替代物的功能性变异频率时，很多研究将面对相关基因中的给定 LD 功能块与感兴趣疾病之间的相关性力不从心（Ke et al. 2004）。

在实际工作中，研究者尝试利用 tag-SNP 和单体型作为代用标记来代替 LD 功能块中潜在的功能变异。代用标记的性能与代用标记与通过 $R^2$ 测量的功能变异间的相关性

成正比（Ke et al. 2004）。不幸的是，如果代用标记与功能变异间的频率有一定差距，病例/对照样本联合研究的能力就会有本质的丧失。例如，如果某研究者利用一个群体等位基因频率为 30%的普通单体型或 tag-SNP 来检测频率为 10%的功能变异，其性能将有 10 折的落差。这样，对于针对 300 个患者和 300 个对照个体的病例/对照样本联合分析来说，其性能只相当于 30 个患者和 30 个对照个体的研究，昂贵的样本搜集和辛苦的基因型分析劳动大都付诸东流。

因此，在研究中，无论是确定每个 LD 功能块中的单体型涵盖范围还是单体型多样性，谨慎的操作都很必要。产生一个单体型多样性量值是有益的。如果我们只在获得功能变异的有效替代品，假定我们要达到在一个一般群体中获得低到 5%～10%的等位基因频率的变异的力度，那么一组理想的单体型的数量将是 10 个，每一个具有等值的10%等位基因频率。所以，我们干脆将单体型多样性的测算方法确定为单体型频率的总和，其中每个单体型都小于或等于 10%。例如，如果存在着每个单体型的频率为 20%的 3 个普通单体型和每个单体型的频率为 20%的 4 个其他单体型，单体型多样性将等于 40%，我们说 60%的单体型保留空间不具备足够的多样性。

在平均为 60 kb 的 LD 功能块中，可以选择 6～10 个普通 tag-SNP 来考虑 50～60个 HapMap SNP 中的大量信息。这可以获得基因型分析花费的实质性储备。假定历史上在标记之间没有重组发生，那么预期的情况将是在一个 LD 功能块中由 N 个 SNP 标记的独立单体型不会多于 N＋1。这样，能够确定的处于一个强大的 LD 功能块中的tag-SNP 只有 6 个，至多 7 个单体型（每个单体型中一个变异）。然而，7 个独立单体型将具有 14%的平均频率，这不能满足愿望中的理想标准，即每个单体型的频率均不超过 10%。解决这一问题的一个方法是，进一步在每一个感兴趣的 LD 功能块中加上一个或两个微卫星以将普通 tag-SNP 确定的单体型分开。

在分析疾病的等位基因相关性时，微卫星补充了 SNP 作为有用标记的不足，尽管在群体历史上它们不如 SNP 那样稳定。不过，它们的平均突变率比 Weber 原来报告的要低些（表 18-1）（Beckman and Weber 1992）。基于首次由证明符合孟德尔规律的成千上万个微卫星所做的标记，采用 30 个三组一套的平板试验，我们的估计是突变率为大约 $5\times10^{-4}$或每 2000 代一个突变。这与其他两项研究是一致的，它们针对一个很大数量的微卫星（Huang et al. 2002；Whittaker et al. 2003）。这一突变率对于维持替代标记和一个高频率的潜在功能性疾病变异间的强大 LD 是很不理想的。然而，这一突变率对于中等频率（如 2%～40%）的普通功能变异实际上是理想的。作为一个例子，微卫星可以标记中等频率的 HLA 扩展单体型（Vorechovsky et al. 2001）。因为微卫星可能具有 5～20 个变异，它可以在一个给定的 LD 功能块中有效地分开单体型空间。进一步，微卫星还可以将共同祖先的 SNP 单体型分开，使其每份还有较少的微卫星/单体型。这提供了分析任何给定的 LD 功能块中大量增加的单体型多样性的有效方法。举一个真实的例子，表 18-2A 显示了 10 个通过大约 100 000 个碱基的单一 LD 功能块中的 6个 tag-SNP 确定的独立单体型。大约 70%的单体型空间不足于涵盖 10%频率或等频率的单体型。表 18-2B 中只加上 1 个微卫星的情况就导致普通单体型的多样化。现在，没有哪个 SNP/微卫星单体型的频率高于 10%。例如，表 18-2A 中第一个 SNP 单体型具

有一个高频率，但通过添加微卫星就分开成为至少两个低频单体型（表 18-2B）。注意，微卫星也可能过分变化不定，使得单体型分裂为太多的低频单体型，其频率可能低于功能变异。不过，单体型相关分析可以通过具有微卫星的两轮 SNP 分析和没有微卫星、而具有对两倍试验数目的合适的 p-值校正值的 SNP 分析来完成。

**表 18-1　每代微卫星突变率的先前估计**

| 每代突变率 | 参考文献 |
|---|---|
| 0.000 194 | Huang 等.（2002） |
| 0.000 45 | Whittaker 等.（2003） |
| 0.000 56 | deCODE 的经验 |

**表 18-2　一个给定的 LD 功能块中的单体型多样性**

| A. tag-SNPs only | | | | | | | | | B. Addition of 1 microsatellite | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| R1438 | RS7627220 | RS7651931 | RS4894808 | RS2124546 | RS6787115 | # Markers | # | Freq. | RS1438 | RS7627220 | D3S3725 | RS7651931 | RS4894808 | RS2124546 | RS6787115 | # Markers | # | Ferq. |
| 3 | 2 | 1 | 2 | 4 | 3 | 6 | 60 | 0.321987 | 3 | 2 | 12 | 1 | 2 | 4 | 3 | 7 | 60 | 0.090376 |
| 1 | 1 | 3 | 3 | 2 | 1 | 6 | 60 | 0.161569 | 3 | 2 | 14 | 1 | 2 | 4 | 3 | 7 | 60 | 0.077723 |
| 1 | 1 | 3 | 3 | 4 | 3 | 6 | 60 | 0.130702 | 1 | 1 | 0 | 3 | 3 | 4 | 3 | 7 | 60 | 0.068971 |
| 3 | 2 | 1 | 3 | 4 | 3 | 6 | 60 | 0.108933 | 3 | 1 | 4 | 1 | 3 | 4 | 3 | 7 | 60 | 0.06465 |
| 3 | 1 | 1 | 3 | 4 | 3 | 6 | 60 | 0.064655 | 1 | 1 | 6 | 3 | 3 | 4 | 3 | 7 | 60 | 0.063353 |
| 3 | 2 | 1 | 3 | 2 | 1 | 6 | 60 | 0.054627 | 3 | 2 | 12 | 1 | 3 | 4 | 3 | 7 | 60 | 0.062264 |
| 3 | 2 | 1 | 2 | 2 | 1 | 6 | 60 | 0.035271 | 3 | 2 | 10 | 1 | 2 | 4 | 3 | 7 | 60 | 0.05929 |
| 3 | 2 | 1 | 3 | 4 | 1 | 6 | 60 | 0.034676 | 1 | 1 | 6 | 3 | 3 | 2 | 1 | 7 | 60 | 0.047222 |
| 3 | 1 | 3 | 3 | 2 | 1 | 6 | 60 | 0.034146 | 3 | 2 | 8 | 1 | 2 | 4 | 3 | 7 | 60 | 0.044808 |
| 1 | 1 | 1 | 3 | 4 | 3 | 6 | 60 | 0.009132 | 1 | 1 | 0 | 3 | 3 | 2 | 1 | 7 | 60 | 0.042155 |

A. the SNPs only（仅有 SNPs）；B. addition of 1 microsatellite（加上 1 个微卫星）。

# Markers：标记；Freq.：频率。

在实际中，微卫星在检测普通疾病的显著关联性时用处很大。例如，2 型糖尿病的广泛复制基因 TCF7L2 就是通过对包含 1200 个微卫星的冰岛家族的全基因组连锁扫描进行定向克隆而得以发现的（Reynisdotteir et al. 2003）。对冰岛的 1185 个患者和 931 个对照个体的持续的基因座位范围的病例对病例相关分析应用了 228 个微卫星标记，针对 two-LOD（差别的 log 值）的研究，使第 10 号染色体上显示的连锁峰下降了一个间隔（10.5 Mb）（Grand et al. 2006）。关联研究（association study）所要求的微卫星空间平均大小曾经为 46 kb，其结果是每个 LD 功能块具有 1～2 个微卫星。在利用 Fisher 法精确测试的单标记关联中，发现了一个单一的微卫星标记 DG10S478，位置处于 TCF7L2 基因的第三个内含子内，它与最普通的等位基因显著关联，相对风险值（RR）等于 0.67（未校正的 $p$ 值为 $2.1\times10^{-9}$）。我们注意到，其他频率高于 2%的等位基因

体现的 RR 值为 1.21～1.53［表 18-3（i）］。我们将所有非普通等位基因合并成一个复合型普通等位基因，取个名字 X，它给出一个显著的 RR 值 1.5 和一个大小为 0.276 的群体等位基因频率。研究发现，微卫星关联在一个丹麦群（cohort）和一个美国群中发生了复制［表 18-3（ii）］。注意，在这 3 个高加索群体之间等位基因频率和 RR 值是非常相似的。RR 总值是 1.56，其校正的 $p$ 值是 $7.8\times10^{-15}$；在一个乘法模型（Grant et al. 2006）中纯合子对应的 RR 值为 2.4。

**表 18-3　微卫星关联研究**

| | 等位基因 | 受影响频率 | 可控制频率 | RR 值 | P 值 |
|---|---|---|---|---|---|
| (i) DG10S478 所有等位基因 | | | | | |
| 冰岛（1185/931） | 0 | 0.636 | 0.724 | 0.67 | $2.1\times10^{-9}$ |
| | 4 | 0.005 | 0.002 | 2.36 | 0.12 |
| | 8 | 0.093 | 0.078 | 1.21 | 0.090 |
| | 12 | 0.242 | 0.178 | 1.48 | $4.6\times10^{-7}$ |
| | 16 | 0.022 | 0.015 | 1.53 | 0.076 |
| | 20 | 0.001 | 0.003 | 0.39 | 0.17 |
| (ii) DG10S478 的等位基因 X | | | | | |
| 冰岛（1185/931） | x | 0.364 | 0.276 | 1.50［1.31，1.71］ | $2.1\times10^{-9}$ |
| 丹麦（228/539） | x | 0.331 | 0.260 | 1.41［1.11，1.79］ | 0.0048 |
| 美国（361/530） | x | 0.385 | 0.253 | 1.85［1.51，2.27］ | $3.3\times10^{-9}$ |
| 总计 | x | — | — | 1.56［1.41，1.73］ | $4.7\times10^{-18}$ |

在我们复制了这一标记后，我们在 TCF7L2 中相同的 LD 功能块内寻找可能具有等价或更好的风险值的 HapMap SNP。有 5 个 SNP 显示出具有普通微卫星等位基因的较强的 LD，并且都与 2 型糖尿病有显著的联系，但是都不如微卫星那样被证明有强大的关联性或者 RR 值（Grant et al. 2006）。最好的 SNP rs7903146 现在已经在大约 20 个群体内被复制，包括高加索人（如 Damcott et al. 2006；Florez et al. 2006；Groves et al. 2006）和非洲人（Helgason et al. 2007）。我们最近的基因组范围关联（GWA）扫描利用的是来自一个 317 K Illumina 珠子芯片阵列的基因型分析，这些分析针对 1491 个冰岛 T2DM 患者和 4712 个对照个体。这些分析发现，具有最好的单标记关联的 SNP 与 TCF7L2 中的 SNP rs7903146 相同。上文已经介绍过，rs7903146 是通过定向克隆发现的。这个关联具有一个未校正的 $p$ 值 $3.7\times10^{-13}$，假定一个多重试验模型，远远超过阈值 $1.3\times10^{-7}$。

微卫星标记应用于串联研究的第二个例子是在前列腺癌的染色体 8q 上的广泛推广。在进行串联研究时，利用其全基因组的 1200 个微卫星标记在染色体 8q 上定位前列腺癌基因（Amundadottir et al. 2006）。在一个 10 Mb 的片段上，在密度为 28 kb 的连锁高峰处，358 个微卫星被设计用于患者/对照群的基因型检测，用了 869 例没有亲缘关系的冰岛患者和 596 例对照。研究发现了 DG8S737 的普通等位基因 0 的相关性，其 RR

为 1.79——即使考虑到测试的数量，这也是很显著的。这个微卫星关联与前列腺癌的相关性在瑞典、美国和高加索前列腺癌患者群中有拷贝。然而，最有趣的方面是，这个微卫星标记在非裔美国人中比在高加索人中更强大［人口隶属风险度（population-attributable risk）为 16%比 8%］。这表明，微卫星可用于研究人口，甚至跨越种族人口。相同的微卫星标记已被复制到其他人群（Freedman et al. 2006）。我们最近使用来自 Illumia 的 317K SNP 珠子芯片，在 GWA 研究中再一次显示出与一个到达这一微卫星的强 LD 区的 SNP 的最显著结合。根据一个人的观察角度，这些糖尿病和前列腺癌的研究，验证了以下 4 种方式中的一个或多个：①连锁关系对于定位普通的疾病基因是有用的，②GWA 对于普通的疾病基因的发现是有用的，③微卫星对于常见的疾病基因和它们的复制是有用，④Illumina 317K 珠子芯片产生的数据，似乎不受产量差距或等位基因在病例组和对照之间如何被命名的影响（这可能会导致巨大的假阳性信号，淹没了真正的阳性）。

## 病例组/对照组研究中应用微卫星的实验方法

本节对于实践中在候选区域或基因内对病例组/对照组关联进行研究提供更详细的微卫星使用信息。在候选基因的研究中，微卫星标记对每个 LD 功能块选定的 tag-SNP 都是有益的补充。关于每个与候选基因 LD 功能块中 tag-SNP 关系的典型研究，利用 SNP 选择程序（如 Tagger）（de Bakker et al. 2006）在与候选基因重叠的每个 LD 功能块中选择 5 个 SNP 标记。一个或两个微卫星标记也可以被添加到每个 LD 功能块中，使单体型空间多样化。

微卫星标记可能是从 Benson 开始用串联重复探测者（TRF）程序发现的（1999）。以充足的标准设计的引物被用于 PCR 试验，而包括重复序列或已知 SNP 的引物被排除在外。我们设置了一套先前设计的并在 deCODE 系统中测试过的超过 25000 个的微卫星标记进行实验，获得成功。随后，我们通过将微卫星的特点与这些成功经验相联系来估计一个微卫星标记成功（即可读的和多态性标记）的可能性，包括单体长度、微卫星长度和重复结构。我们运用这一信息对通过预期质量推测的微卫星进行定级，并已经在基因组中确定了超过 40 000 个推测的微卫星，将其归类为 1 级。根据我们的经验，超过 70%的这些微卫星标记被证明是可用的。然而，相当一部分的低级别预测微卫星也能带来有效的分析结果，因此，在候选区域内 10～40 kb 的密度通常是能够实现的。

在跟踪随访病例/对照关联以筛选疾病基因的研究过程中，我们发现微卫星标记对于超细连锁高峰间隔是有用的。虽然对于显著的基因组范围的连锁关系，人们只是选择那些符合标准的连锁高峰，但是我们发现，即使是潜在的连锁高峰，对于挑选基因区域的超细定位也是有用的。事实上，对 2 型糖尿病和前列腺癌的研究表明，迄今为止在导致这些疾病的大多数广泛复制的基因中，这个连锁峰只是中等水平的。然而，重要的是，我们只应用一定的影响因素，对没有遗传模式特殊性的等位基因进行共享连锁分析。这大大降低了测试的数目，因为一个参数连锁研究可能着眼于多于 100 个可导致较高假阳性比率的连锁分析模型。

我们还发现在高密度 GWA 研究中微卫星标记可用于追踪最强的关联信号。在

GWA 研究中大多数最强关联是基于风险的和相当适中的（如 1.2～1.4）很普通的等位基因。它对追踪一些信号是有用的。例如，在发现人群中同一 LD 功能（以及在同一基因的其他 LD 功能块）中使用额外标记的信号，包括额外的 SNP 标签及一个或两个微卫星。这可能使得原始信号能被更好地捕捉，但 RR 较高，或有可能导致在同 LD 功能块或相同基因的其他区域发现其他疾病相关变异。如果人们在芯片上发现比原来的 tag-SNP 标记更强的信号，如果它是一个真正的疾病基因，就可以提高复制的机会。

使用微卫星标记进行全基因组连锁研究，这个方法得到的数据在目前 GWA 时代仍然是有用的，尤其是当 P 值必须通过 317 000 或 550 000 SNP 标记实验纠正时。人们可以实现连锁间隔的预设，这被用在 GWA 研究的原始分析中。例如，如果选择只覆盖基因组 3%最高的 5 个或 6 个潜在的连锁峰，此后的纠正可能只针对这些峰的 10 000 SNP，而不是所有 317 000 个试验。二级分析可以涉及基因组其余的部分。

对于最强烈的连锁峰，来自珠子芯片的高密度 SNP 基因型分析至少对于追踪病例/对照关联的连锁间隔的超细定位是非常有效的方法。然而，通过微卫星对如此高密度的数据进行补充以增加单体型多样性，是有用的，因为平均每个 LD 功能块只有 6～9 个 tag-SNP 标签。

确定单体型多样性是一种用来确定每个候选基因或间隔 LD 功能块的覆盖是否已经足够的方法。如上所述，一个目标是利用大于 10%的足够的 SNP 和微卫星标记某一 LD 功能块而不留下任何单体型。一个 LD 功能块不足的单体型多样性可能导致疾病相关的功能变异处于一个非常普遍的祖先单体型的背景上。如果变异只在普通单体型的一部分，那么单体型可能不会显示与疾病有足够紧密的联系，同时基因关联将被忽略。在这种情况下，单体型不是一个潜在的功能变异的良好替代物。将微卫星标记整合到 LD 类型之中，是一件很简单的事。通过确定由边缘等位基因概率衡量的两个标记的所有可能的等位基因组合的平均值，$D'$和 $R^2$ 的标准定义得以扩展，包括微卫星标记在内。比较患者人群和对照人群，单点和单体型关联按照本手册的其他章节所描述的程序进行研究。虽然没有公认的标准来界定单体型或 SNP 关联的统计学显著性，但很显然，$p$ 值应该通过测试数目加以纠正［包括标记的数目（等位基因数-1)、表型的类别数和所用的统计方法］。同样清楚的是，在其他独立的疾病群中的复制也是必要的。

对于任何重大的研究成果，我们必须检查人口的分层问题。重要的是要确认在自我报告中每个群体的种族划分，并确保患者组与对照组的遗传背景相匹配。微卫星标记的设定尤其需要区分主要种族信息。为了评价研究人群以遗传学意义估计的祖先基因，我们使用来自 2000 个微卫星基因型的一组 75 个没有联系的微卫星标记，这个基因型是来自多民族群体的，包括 35 个来自巴尔的摩的欧裔美国人，88 个来自巴尔的摩的匹兹堡和北卡罗来纳州的非裔美国人，34 个中国人（广东）和 29 个美国印第安人（萨波特克）(Helgadottir et al. 2006)。在这 2000 个微卫星标记中，最终选中在欧裔美国人、非裔美国人、亚洲人和印度人中有最显著性差异、同时也有良好的质量和产量的一套。另一项由 Tang 等（2005）承担的研究也利用了这些标记中的 31 个。

利用来自这些具有民族标志样本平板的基因型数据（高加索、西非和亚洲板块，每个平板使用 96 个样品)，结构软件将对个人的遗传背景进行估测（Pritchard et al.

2000；falush et al. 2003）。结构分析推断，K祖先种群的等位基因频率以一套来自个人的位点基因型的和使用者特定的K值为基础，并通过推断K人群的每个分配祖先的比例推断每一个人分配祖先的比例。你可以在两个方面使用种族样品组。首先，可以只分析那些至少有90%欧洲血统的高加索患者和对照组。其次，我们修改了NEMO单体型关联算法，以适应不同遗传背景的患者和对照组，根据自我报告使你可以在分析中包括所有的高加索人（Helgadottir et al. 2006）

# 结论

总之，在连锁和关联这两个研究中微卫星被证明是非常有用的标记。事实上，它们比单碱基替换具有更高的突变率，这使得它们将明显的非突变普通SNP单体分解到较低的频率，这可能更符合其中打算要检测中度频率的或罕见功能的变体。

# 致谢

作者感谢Struan Grand，在表18-2中创建了他在deCODE工作期间获得的单体型多样性数据。

# 参考文献

Amundadottir L.T., Sulem P., Gudmundsson J., Helgason A., Baker A., Agnarsson B.A., Sigurdsson A., Benediktsdottir K.R., Cazier J.B., Sainz J., et al. 2006. A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38:** 652–658.

Beckman J.S. and Weber J.L. 1992. Survey of human and rat microsatellites. *Genomics.* **12:** 627–631.

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.

Damcott C.M., Pollin T.I., Reinhart L.J., Ott S.H., Shen H., Silver K.D., Mitchell B.D., and Shuldiner A.R. 2006. Polymorphisms in the transcription factor 7-like 2 (*TCF7L2*) gene are associated with type 2 diabetes in the Amish: Replication and evidence for a role in both insulin secretion and insulin resistance. *Diabetes* **55:** 2654–2659.

Davies J.L., Kawaguchi Y., Bennett S.T., Copeman J.B., Cordell H.J., Pritchard L.E., Reed P.W., Gough S.C., Jenkins S.C., Palmer S.M., et al. 1994. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371:** 130–136.

de Bakker P.I., Burtt N.P., Graham R.R., Guiducci C., Yelensky R., Drake J.A., Bersaglieri T., Penney K.L., Butler J., Young S., et al. 2006. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38:** 1298–1303.

Falush D., Stephens M., and Pritchard J.K. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164:** 1567–1587.

Florez J.C., Jablonski K.A., Bayley N., Pollin T.I., de Bakker P.I., Shuldiner A.R., Knowler W.C., Nathan D.M., and Altshuler D. (Diabetes Prevention Program Research Group). 2006. *TCF7L2* polymorphisms and progression to diabetes in the Diabetes Prevention Program. *N. Engl. J. Med.* **355:** 241–250.

Freedman M.L., Haiman C.A., Patterson N., McDonald G.J., Tandon A., Waliszewska A., Penney K., Steen R.G., Ardlie K., John E.M., et al. 2006. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci.* **103:** 14068–14073.

Grant S.F., Thorleifsson G., Reynisdottir I., Benediktsson R., Manolescu A., Sainz J., Helgason A., Stefansson H., Emilsson V., Helgadottir A., et al. 2006. Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat. Genet.* **38:** 320–323.

Gretarsdottir S., Sveinbjornsdottir S., Jonsson H.H., Jakobsson F., Einarsdottir E., Agnarsson U., Shkolny D., Einarsson G., Gudjonsdottir H.M., Valdimarsson F.M., et al. 2002. Localization of a susceptibility gene for common forms of stroke to 5q12. *Am. J. Hum. Genet.* **70:** 593–603.

Groves C.J., Zeggini E., Minton J., Frayling T.M., Weedon M.N., Rayner N.W., Hitman G.A., Walker M., Wiltshire S., Hattersley A.T., and McCarthy M.I. 2006. Association analysis of 6,736 U.K. subjects provides replication and confirms *TCF7L2* as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* **55:** 2640–2644.

Gulcher J. and Stefansson K. 2006. Positional cloning: Complex cardiovascular traits. *Methods Mol. Med.* **128:** 137–152.

Helgadottir A., Manolescu A., Helgason A., Thorleifsson G., Thorsteinsdottir U., Gudbjartsson D.F., Gretarsdottir S., Magnusson K.P., Gudmundsson G., Hicks A., et al. 2006. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat. Genet.* **38:** 68–74.

Helgason A., Palsson S., Thorleifsson G., Grant S.F., Emilsson V., Gunnarsdottir S., Adeyemo A., Chen Y., Chen G., Reynisdottir I., et al. 2007. Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39:** 218–225.

Huang Q.Y., Xu F.H., Shen H., Deng Y.J., Liu Y.J., Liu Y.Z., Li J.L., Recker R.R., and Deng H.W. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. *Am. J. Hum. Genet.* **70:** 625–634.

Ke X., Durrant C., Morris A.P., Hunt S., Bentley D.R., Deloukas P., and Cardon L.R. 2004. Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* **13:** 2557–2565.

Kong A., Gudbjartsson DF., Sainz J., Jonsdottir G.M., Gudjonsson S.A., Richardsson B., Sigurdardottir S., Barnard J., Hallbeck B., Masson G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.

Murray S.S., Oliphant A., Shen R., McBride C., Steeke R.J., Shannon S.G., Rubano T., Kermani B.G., Fan J.B., Chee M.S., and Hansen MS. 2004. A highly informative SNP linkage panel for human genetic studies. *Nat. Methods* **1:** 113–117.

Palsson B., Palsson F., Perlin M., Gudbjartsson H., Stefansson K., and Gulcher J. 1991. Using quality measures to facilitate allele calling in high-throughput genotyping. *Genome Res.* **9:** 1002–1012.

Pritchard J.K., Stephens M., and Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

Reynisdottir I., Thorleifsson G., Benediktsson R., Sigurdsson G., Emilsson V., Einarsdottir A.S., Hjorleifsdottir E.E., Orlygsdottir G.T., Bjornsdottir G.T., Saemundsdottir J., et al. 2003. Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. *Am. J. Hum. Genet.* **73:** 323–335.

Tang H., Quertermous T., Rodriguez B., Kardia S.L., Zhu X., Brown A., Pankow J.S., Province M.A., Hunt S.C., Boerwinkle E., Schork N.J., and Risch N.J. 2005. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am. J. Hum. Genet.* **76:** 268–275.

Vorechovsky I., Kralovicova J., Laycock M.D., Webster A.D., Marsh S.G., Madrigal A., and Hammarstrom L. 2001. Short tandem repeat (STR) haplotypes in HLA: An integrated 50-kb STR/linkage disequilibrium/gene map between the RING3 and HLA-B genes and identification of STR haplotype diversification in the class III region. *Eur. J. Hum. Genet.* **9:** 590–598.

Weber J.L. and Broman K.W. 2001. Genotyping for human whole-genome scans: Past, present, and future. *Adv. Genet.* **42:** 77–96.

Whittaker J.C., Harbord R.M., Boxall N., Mackay I., Dawson G., and Sibly R.M. 2003. Likelihood-based estimation of microsatellite mutation rates. *Genetics* **164:** 781–787.

# 19 SNP 选择时的考虑事项

Chris Carlson

*Fred Hutchinson Cancer Research Center, Seattle, Washington 98109-1024*



## 简介

人类基因组由大约 30 亿个碱基对的核苷酸组成（Lander et al. 2001；Venter et al. 2001）。当比较任意两个基因组序列拷贝时，绝大多数序列是保持一致的，大约每 1000 个碱基对的序列中存在 1 个序列变异。这些变异中绝大多数以 SNP 的形式存在，单核苷酸的替换能大约占 SNP94%，1～4 个碱基的缺失能占其中 4.5%，其余部分为 1～4 个碱基的插入变异（Bhangale et al. 2005）。其他类型的序列变异，还包括微卫星重复、多等位基因的可变数目串联重复，转座子插入以及更大的结构多态性变异（插入/缺失和转位）。本章节中，绝大部分主要围绕 SNP 进行介绍，由于其能解释个体间绝大多数的等位基因变异。

## tag-SNP 选择的理论背景和目的

多年以来，识别功能性重要的 SNP 一直是人类遗传研究领域的重要研究领域。这个领域的研究随着时间的推移以及不断的科学努力，也取得了逐步的进展。最容易识别的功能性变异是能产生显著效应的罕见变异，其经常与孟德尔遗传的疾病相关联。最初，这些变异的识别主要来自候选基因分析（如镰刀细胞贫血），通过连锁分析可以使得遗传学家筛选基因组以寻找包含这些主效基因的基因组区段。许多连锁区段通过候选基因分析都成功地缩窄得到点突变的变异结果，但是，每个连锁区段中的真正效应基因有相当一部分直至基因组计划完成也没有被识别出来。甚至现在，并不是所有的连锁区段都包含明显的候选基因，并且明显的候选基因也往往并非都包含功能性的变异位点。然而，连锁作图和定位克隆技术在识别与疾病高发生率相关的罕见变异方面已经取得了成功。

连锁分析主要在罕见疾病的遗传学领域取得了成功。甚至有时候在常见疾病领域中，对于罕见和表型高外显率的常见疾病的遗传定位，连锁分析也能取得成功（如 BRCA1）虽然在特定家系中，相关的变异位点能对很大一部分的疾病风险进行解释，但是对于全人群中该疾病的整体发生，这些罕见的变异位点却只能解释其中相对很少的一部分。因而，虽然许多疾病的发生过程中遗传因素占据了明确的地位，但是只有很少一部分遗传病因能被主效的罕见变异所解释。某个等位基因的遗传相对危险度（GRR）可以被定义为个体携带的单个等位基因拷贝所对应的疾病风险的改变。理论分析表明，连锁分析可能不能识别低于 4 倍的 GRR（Risch and Merikangas 1996）。

关联分析和连锁分析有所不同，其主要用于无关病例和对照人群的分析，其目的在于识别等位基因在两组人群中分布频率的显著性差异。尽管连锁分析能筛选同一家系中受累的个体共有的基因组区段，但是关联分析却能筛选“无关”个体人群共有的风险等位基因。因为大多数人类基因组中的 SNP 等位基因可以追踪为单独的突变事件（大多数 SNP 未经历再发突变），这使得筛选无关个体人群以寻找来自古代共同祖先共有的基因组片段。虽然染色体大片段（兆碱基对）通常在单一家系中的受累个体中所共有，但是，由于重组事件的较少次数，无关个体间共有的染色体区段相对较短（千碱基对）。

由于来自祖先的风险关联 SNP 的染色体侧翼序列片段很短，因而需要相对更多的分子标记在关联研究中以识别这些片段。但是，对于识别常见微效变异，关联分析可能比连锁分析更为有效（Risch and Merikangas 1996）。关于常见微效变异可能比罕见微效变异是否更为普遍还存在争议（Lander et al. 2001）。但是，对于一些孟德尔疾病，大多数受累的家系中共有的风险等位基因相同（如镰刀红细胞贫血和囊肿性纤维化），然而某些疾病风险等位基因（如 BRCA1/2），甚至某些风险基因位点（如 MODY 的风险基因位点）对于其他疾病而言却是不致病的。因而，虽然疾病较为罕见，但是一些孟德尔疾病却大部分由单独的“常见”等位基因引发（囊肿性纤维化，镰刀红细胞贫血，血色素沉著症），虽然也有一些疾病不是这样（MODY，BRCA1/2，着色性干皮病）。因而，似乎一些微效风险等位基因是常见的，尽管其他一些是罕见的。

如果我们特定研究风险等位基因纯合子的相对危险度（RR）以及疾病风险等位基因的频率，估计其在给定样本量中的检验效能是相对容易的。对等位基因频率在病例和对照组间的频率期望值的真实差异进行估计，并且对该差异在可获得的样本中进行识别的精确度进行估计（图 19-1）。十分显然，较大相对危险度的检验效能要优于较小相对危险度。但是，仔细观察此图可以发现，无论相对危险度大小如何，对于罕见等位基因识别的检验效能都很弱［如次要等位基因频率（MAF）低于 10%或大于 90%］。因此，尽管我们有理由假设微效 RR 罕见变异可能存在，但是即便这样我们对于一般的样本量也没有足够的检验效能识别出这样的变异等位基因。

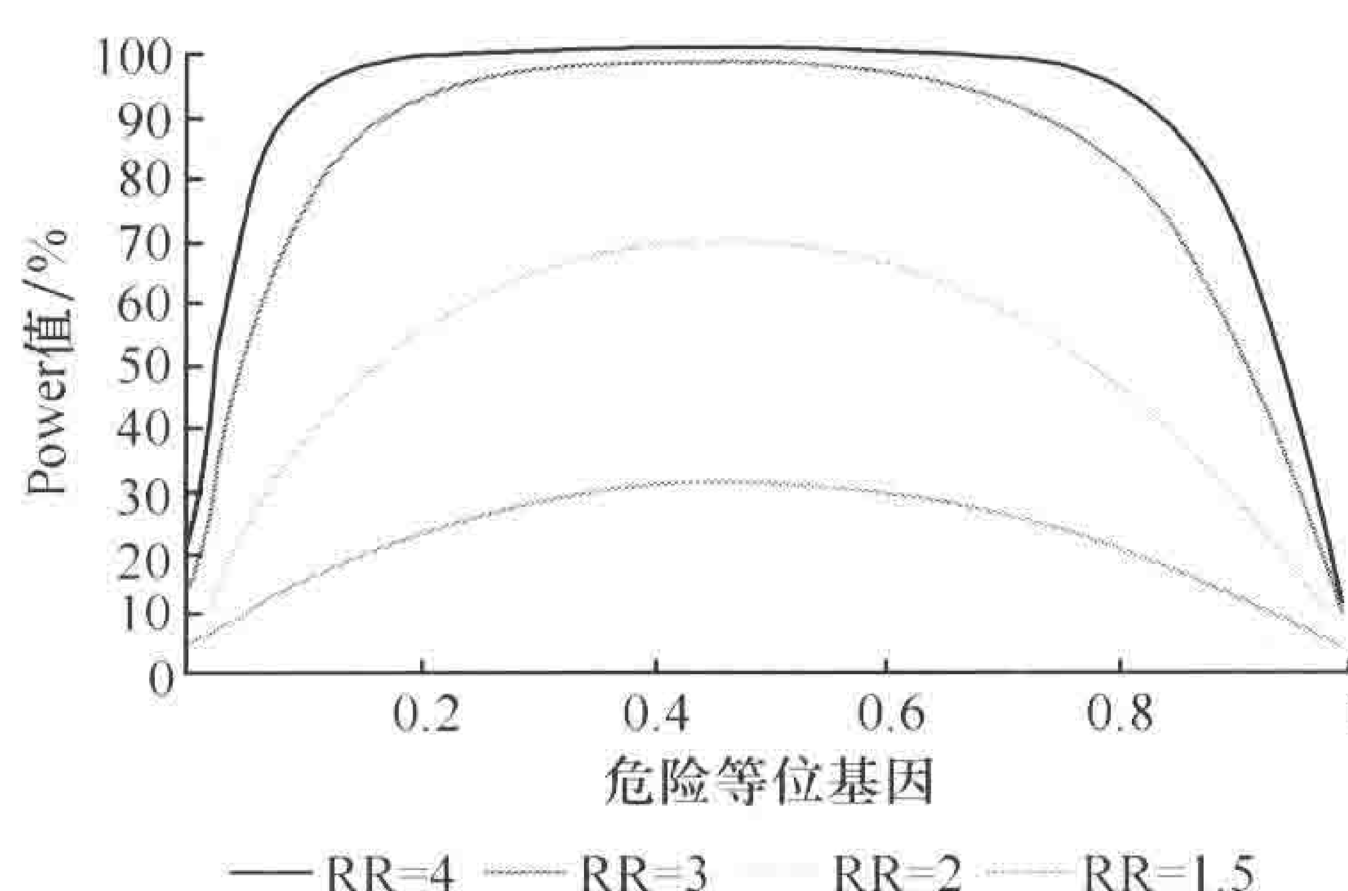



图 19-1　相对危险度的检测分析。200 个病例和 200 个对照组成的人群中 RR 的范围如图所示，采用了风险遗传模型，其中杂合子的 RR 为对应纯合子的 RR 的平方根，没有进行多重检验的调整

举一个简单的例子（Leonid Kruglyak，pers. comm.），假设突变率为每代次 $2\times10^{-8}$/bp，并且每个基因组由 30 亿个碱基对组成，每个基因组在每代次将大约携带 60 个新发突变。假设全球的人口有 60 亿，每个个体有 2 个拷贝的基因组，那么目前每代次将携带 7200 亿个新发变异，或者每个碱基对有泊松均值为 240 个新发变异。因此，基因组中每个碱基对中，均可能有一个与生存一致的杂合新突变，在这个星球上至少一个人的身上发生了。但是，如果我们比较基因组的两个拷贝，我们将会发现 300 万个差异，其中 120 个代表的是新发突变。基因组间大多数差异已经存在了相当久远的时间，并且代表了常见变异的等位基因差异。

前面的分析旨在为关联分析的识别提供支持。关联研究可以用于检验常见疾病/常见变异（CDCV）假设（Collins et al. 1997），但是理论上也不能否认微效罕见变异存在的可能性。我们只是在可行研究的样本量支持下，缺少识别这些变异的检验效能。由于人类基因组的常见 SNP 的数目远少于罕见 SNP 的数目，因而我们可以十分方便的对 CDCV 假设进行检验。估计在人类基因组中存在 600 万～1000 万个常见变异（MAF＞10%）（Kruglyak and Nickerson 2001），重测序的数据结果与所估计的也保持一致。即便允许一些人群特异性的常见变异位点存在，在至少一个主要的地理人群亚群中存在的常见 SNP 变异可能也不会超过 1500 万个。

## tag-SNP 选择的理论方法

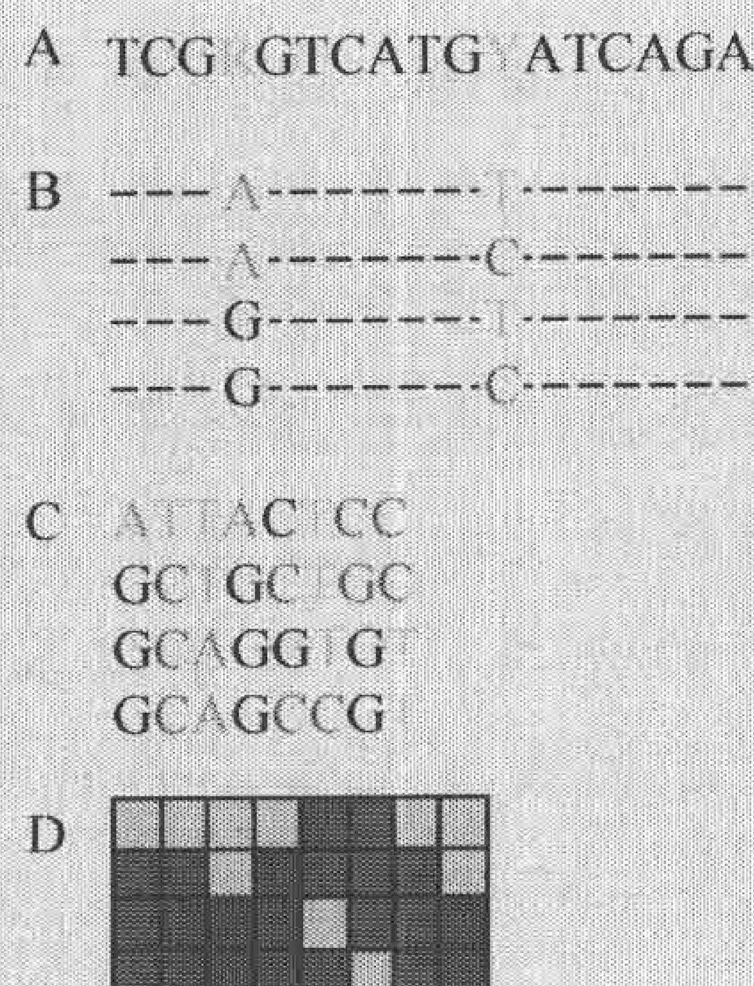
现在，大多数关联研究都针对候选基因、基因组区域或特定信号通路进行，不久的将来全基因组关联分析逐渐将会更频繁的被得到使用。与对于全基因组相比，对于有限的靶序列的范围进行 tag-SNP 的选择相对要容易一些，因而我们将对于特定的基因组靶序列进行研究，但我们可以认识到，仅需要进行很小的调整，相同的方法就可以被应用于全基因组的序列分析之中去。在靶序列区域中，进行关联分析的目的是为了识别功能性的变异。所有的 tag-SNP 策略背后的基础假设是我们不能准确预计哪些 SNP 可能具有功能性的结果。虽然已经有方法可以预测一些类型的功能变异的功能影响。例如，非同义编码 SNP（ncSNP）（Ng and Henikoff 2001，2003），这些 SNP 代表了所有常见变异中的一小部分。因而，选择 tag-SNP 的目的是为了选择一组 SNP，其能够有效描述候选区域中变异的存在模式，以用于后续的基因分型以及关联分析，但并不假设每个 SNP 潜在的功能性的影响作用。

目前有许多种选择 tag-SNP 的算法（Patil et al. 2001；Gabriel et al. 2002；Ke and Cardon 2003；Meng et al. 2003；Sebastiani et al. 2003；Stream et al. 2003；Weale et al. 2003；Zhang and Jin 2003；Carlson et al. 2004；Halldorsson et al. 2004a，b；Lin and Altman 2004；Schulze et al. 2004；Zhang et al. 2004，2005；de Bakker et al. 2005；Halperin et al. 2005.）。在概念上这些算法有许多重叠之处；因此，我们将仅对于几个代表性的算法详细的进行讨论。总的来说，tag-SNP 的选择算法可以被归入两个主要领域：基于单体型的算法和不考虑单体型的算法。基于单体型的算法对于 tag-SNP 选择前，需要推断出单体型的情况（Box19-1，单体型的介绍）。

**Box19-1　单体型的介绍**

单体型是指在同一条染色体上，多个多态位点等位基因的排列模式。大多数单体型分析在使用时需要单体型在小家系中保持稳定。虽然一条完整的染色体可以技术上被认为是一个单体型，但是一般而言单体型是指跨度小于 100kb 且紧密连锁的多态位点，然而也还存在例外（如人类白细胞抗原以及新发突变）。单体型可以通过组合任何多态位点进行构建：可变数目串联重复，微卫星，插入/缺失或者 SNP。

当比较同一物种的任何一对染色体时，会发现其序列的绝大多数都是没有变异的。当论及多态性时，大多数作者采用的阈值为 1%～5%。等位基因沿同一条染色体排列的模式被定义为单体型。如果对于两个复等位基因的多态性位点，理论上就存在 $2^2=4$ 个可能的单体型。随着单体型中包含的多态位点数目的增加，可能的单体型的数目也呈指数增加。对于 $n$ 个复等位基因多态性位点，就存在 $2^n$ 个可能的单体型。理论上，每个碱基的位置上有 4 种可能的核苷酸组成。但实际上，一般来说其仅由两个等位基因组成，因此为了让数据更加直观，许多作者采用常见等位基因和罕见等位基因对单体型进行描述。



A. 多态性的定义是，在基因组的某一位置上存在两种或两种以上的变异，在人群中其出现的频率超过某一规定的阈值（用红色表示）。B. 虽然单体型在技术上包括了非变异位置出现的核苷酸，但是一般来说，仅多态位点的核苷酸将被进行报告。C. 对于附近的多态性位点，观察到的实际单体型的数目往往显著少于理论上可能存在的单体型的数目。D. 对 C 中的数据用图解说明，蓝色和黄色分别代表常见和罕见等位基因

## 基于单体型的 tag-SNP 选择的算法

第一篇描述 tag-SNP 选择（Johnson et al 2001）的文章就采用基于单体型的方法以解决这个问题。他们使用的方法中，根据在基因组的小部分区域中已知存在的少数常见单体型，进行 tag-SNP 的选择以有效的区别于已知的常见单体型。未被作为 tag-SNP 的 SNP 基因型（后文中指的是未分型的 SNP）可以根据个体携带的单体型进行推断从而得知。这种方法对于包含少数常见单体型的基因组区域较为适用，一般来说这对应于很少或不发生重组的区域（Box 19-2，单体型的计算；Box 19-3，非重组区域的单体型，以及 Box 19-4，奠基者事件）。这种“单体型标记”的方法被几个研究单位用于推测更大的基因组区段，这些研究单位认为此算法如果可以用于更大的区域，需要将大区段分割为一系列相对没有重组发生的小区段（“单体型块”），Cardon 和 Abecasis 对这一点进行了评论（2003）。

**Box 19-2　单体型的计算**

单体型是指沿着一条染色体的等位基因排列的模式。因此，单体型也等价于单倍体生物（如细菌）的基因组或染色单体（如雄性哺乳动物的 X 染色体和 Y 染色体）。假设有 $n$ 个无关的双等位基因多态性，理论上将有 $2^n$ 种可能的单体型。在这个分布谱的另一端，非重组的区域，单体型的数目最多为 $n+1$，最少为 2。对于紧密连锁的多态性位点（分布在 0.01cM），存在的单体型数目往往很大程度上少于理论上的最大值。

人类和许多其他模式生物都是二倍体，这样每个个体都携带两套单体型。如果二倍体的个体所携带的由两个或两个以上的多态性位点组成单体型，一条与另一条单体型不同的话，这个个体将被基因分型为多重杂合子，观察到的基因型的模式将与多重单体型一致。对于一个有 $n$ 个多态性位点的杂合子个体，理论上有 $2^{n-1}$ 对可能的单体型。对于未知个体，其单体型的解析可以由许多方法实现。有几种方法可以依靠物理性的分离染色体从而进行单体型的分离，包括等位基因特异性 PCR、亚克隆区域至文库中，以及体细胞杂交，但是这些在这里都不作详细讨论。除此之外，单体型可以根据多种算法利用计算机进行推算，包括家系分析、生物学期望最大化算法以及更复杂的贝叶氏算法。

下面的讨论中，我们使用 IUPAC 模糊代码用于单核苷酸置换的多态性的杂合子基因型的编码：W＝Weak（A 或 T），S＝Strong（G 或 C），K＝Keto（G 或 T），M＝amino（A 或 C），R＝puRine（A 或 G），Y＝pYrimidine（C 或 T）。这样，一个个体如果有三个 SNP 基因型的话，可以表示为 YAR，说明这个个体第一个 SNP 是杂和的 C/T，第二个 SNP 是 A 纯合子，第三个 SNP 是 A/G 杂合子。这些位点在物理距离上不需要靠近；中间的非多态性位点将不会被显示。

基于家系的单体型推算

如果未知个体的父母或后代的基因型已知，那么对于紧密连锁的多态性位点的单体型推算相对较为直接。可以简单的假设，在家系中单体型是非重组的，然后寻找的个体。图 19-2 描述的是一个父母/子女三人核心家系，其中孩子和父亲对于三个 SNP 的基因型（YRR）是多重杂和的。由于母亲是纯合子，那么我们推断这个孩子携带的一个单体型是 CAG，因此，另一个单体型必定就是 TGA。由于这个孩子 TGA 的单体型必须是来自于父亲的，这样也就对父亲的单体型进行了明确。

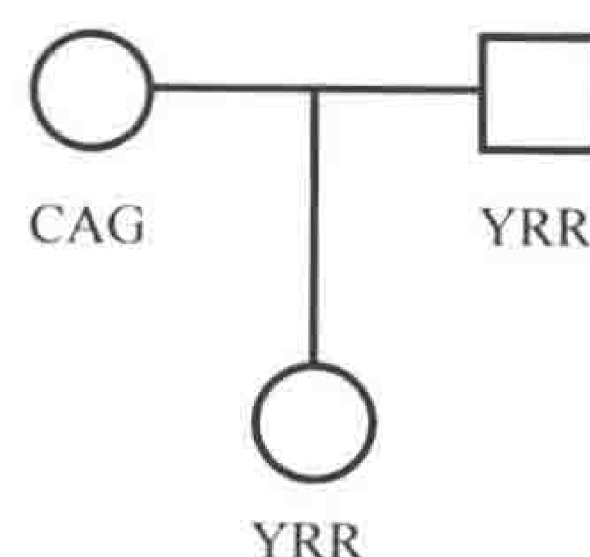
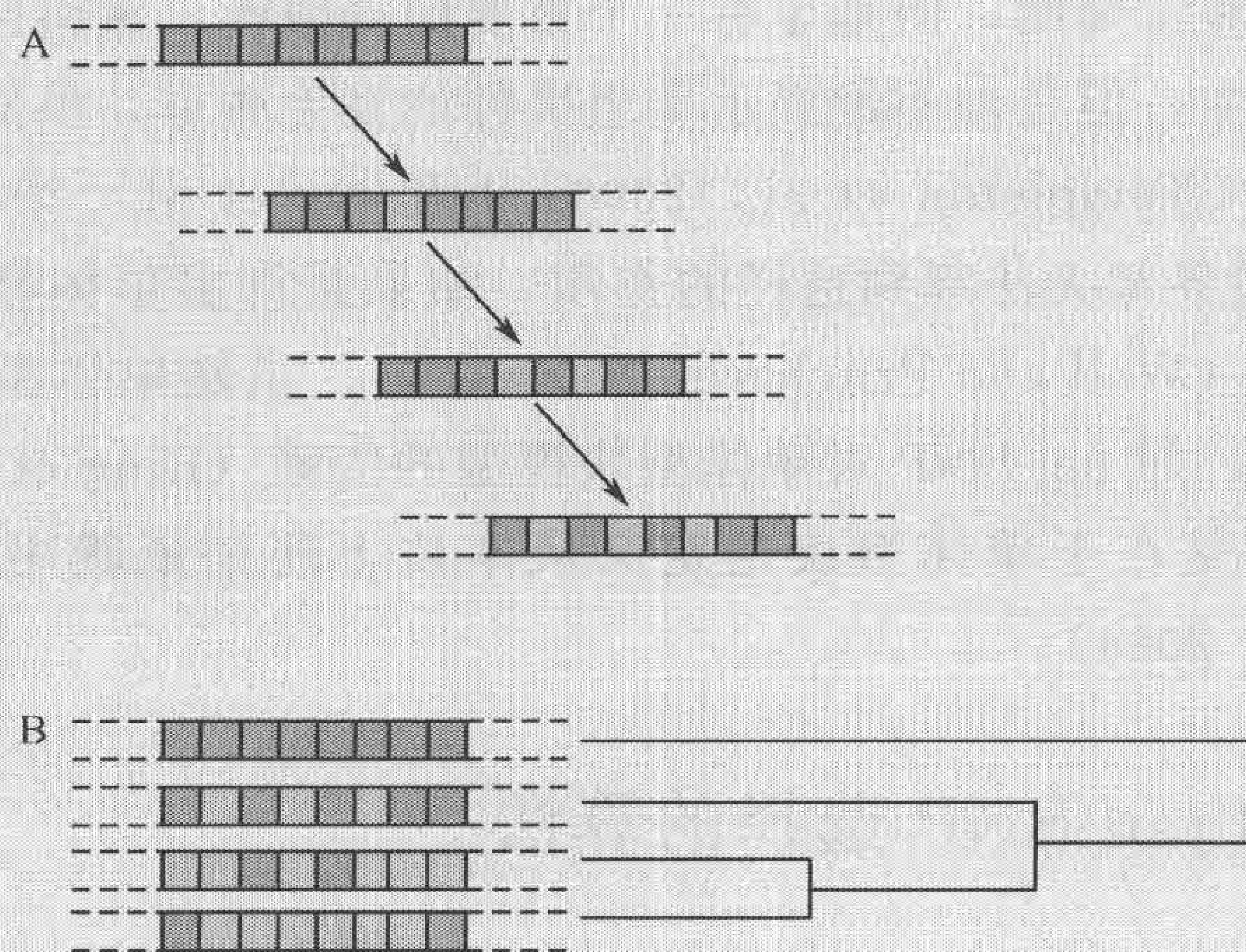



图 19-2　基于家系的单体型推算。父亲（第一排的方块）和孩子（第二排的圆圈）都是三个 SNP 的基因型的多重杂和子（YRR）。因为母亲（第一排的圆圈）是纯合子（CAG），那么孩子携带的其中一条单体型也必须是 CAG。可以推测，其另一条单体型必须是 TGA，那么这条 TGA 的单体型就必定来自于父亲

**Box 19-3　非重组区域的单体型**

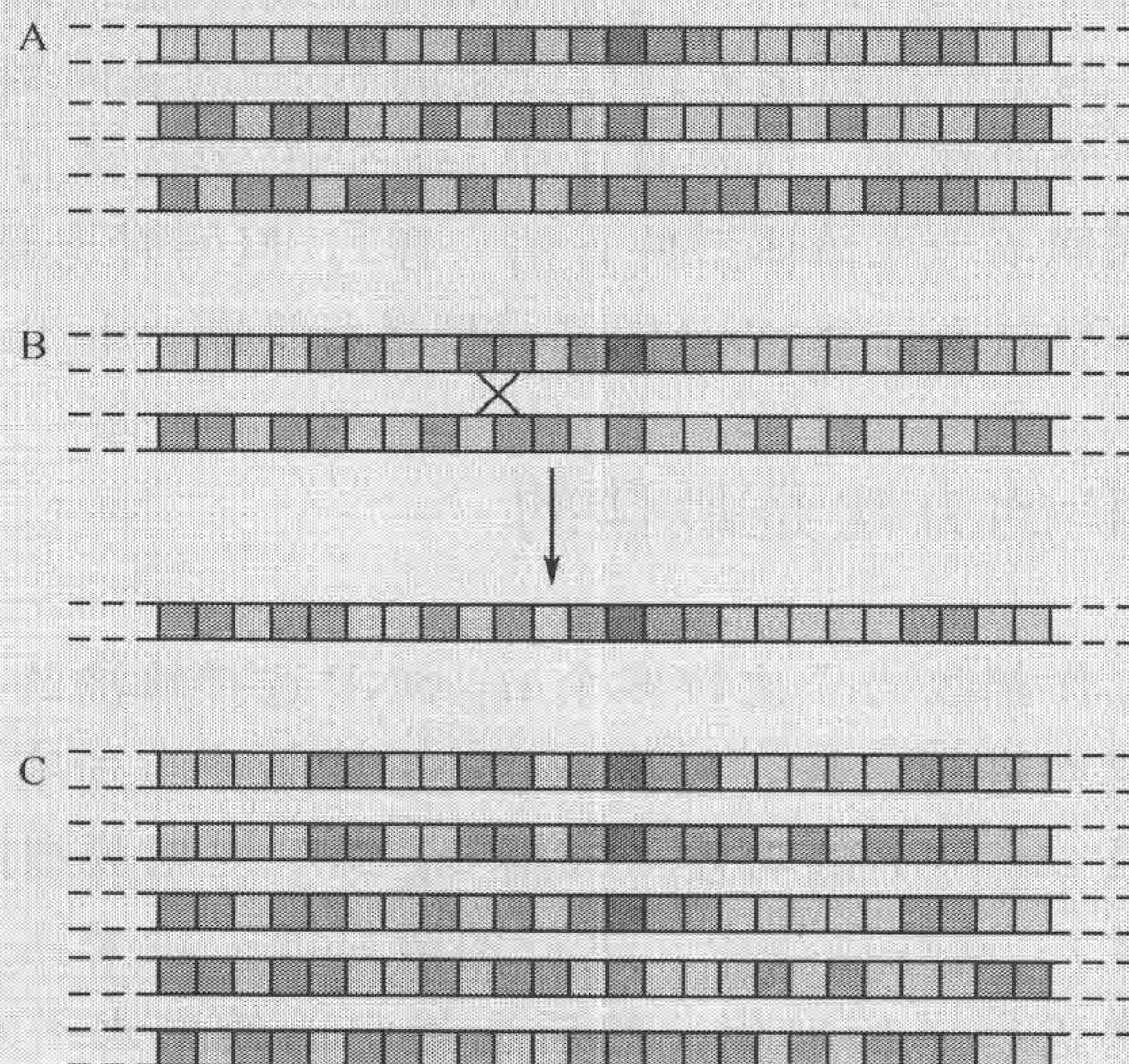
在非重组区域，新单体型主要是由于突变造成的。基因组绝大多数的部分都是非多态性的，因而新的突变常常发生在以前非多态性的位置。这样的话，如果有 $n$ 个多态位点，假设没有再发的突变的话，在非重组区域可能的单体型最大数目为 $n+1$ 个。一般来说，一个人群中的单体型不会包括所有理论上的种类，这是因为其会由于漂变而消失，因而实际的单体型的数目将比 $n+1$ 更少。




A. 在已存在的单体型中出现新的突变，将产生新的单体型。B. 虽然一些中间的单体型不能被观察到，但是根据可观察到的单体型的相似性可以构建单体型关系的进化树

**Box 19-4 奠基者事件**

多态性位点产生的时间与起源单体型和携带的等位基因关联的程度相关。但是，一个等位基因产生的时间也与这个等位基因的频率相关：新的等位基因更少，常见的等位基因则其出现的时间也更久远（Kimura nd Ota 1973；Watterson 1976）。虽然新的突变可能和跨度为兆碱基对的单体型存在关联，但是与常见等位基因（出现的更久远的等位基因）相关联的原始单体型其跨度往往不到 20kb。




A. 在一个已存在的单体型的群体中，在特定的已存在的单体型基础上有一个新出现的突变（红色方块）。B. 由于原始单体型从一代传递至另一代，重组将产生带有原有等位基因的新的单体型。C. 如果在整个群体中，重组带来的等位基因延续下去并有高的频率，那么由于重组的发生，含有该等位基因的单体型所共有的原始单体型的片段在每一代则逐渐减少

单体型块的范例非常简明，但是也有一个重要的局限性：重组区域之间的边界仅能反映重组染色体的发生，但是却不能提示重组事件的发生频率。虽然一些单体型块能清晰的反映重组热点（Templeton et al. 2000；Jeffreys et al. 2001；McVean et al. 2004），但绝大多数边界却无法起到这样的作用，结果是许多单体型块的边界之间也存在很强的连锁不平衡（Wall and Pritchard 2003a，b）。虽然单体型块的模型中还发现了其他的一些局限性，如 tag-SNP 对单体型块模型的依赖（Ding et al. 2005），但是可能最重要的局限性还是在于单体型块的定义从本质上还是依赖用于定义单体型块的 SNP 标记（Ke et al. 2004）。

## 不考虑单体型的 tag-SNP 选择的算法

除了基于单体型的结构进行 tag-SNP 的选择之外，也存在其他许多替代的算法，其依赖于 SNP 之间的相关程度而进行有效的 tag-SNP 的选择（Weale et al. 2003；Calson et al. 2004；Lin and Altman 2004）。这些方法主要是寻找基因组某区域中有效的 tag-SNP 组合，以使得根据对于 tag-SNP 的基因分型可以实现对未分型的 SNP 的基因型进行准确推断。这些方法在算法本质上的优点在于，单体型不需要进行推断，这样可以避免如何对一个单体型块的边界进行定义的问题，同时也可以避免在单体型推断中可能发生的错误。

虽然单体型块的模式是存在局限性的，但是还是存在许多基于单体型信息选择 tag-SNP 的有用的算法。尽管这些算法一般来说并不是最为有效（因为其把临近的“单体型块”作为独立的信息），但是相对于不考虑单体型的方法他们确实更为经济，后者需要根据一系列 tag-SNP 的组合推断出未进行分型的 SNP 的基因型的信息。因此，可能最为有效的方法是混合性的算法，这种算法中 tag-SNP 的选择目的是为了推断未分型 SNP 的基因型信息（如同不考虑单体型的方法一样），但同时也允许根据单个 tag-SNP 或多个 tag-SNP 的单体型对未分型 SNP 的基因型进行推断（de Bakker et al. 2005）。

## 哪种选择 tag-SNP 的方法是最佳的

近来，许多研究工作都致力于比较多个 tag-SNP 选择算法的计算结果的比较。这些算法可以根据许多层面进行比较，包括 tag-SNP 的效率（所有 SNP 中作为 tag-SNP 进行选择的比例）以及 tag-SNP 的功效（根据 tag-SNP 对未分型的 SNP 进行基因型推断等达到如何的准确性）。这些比较中，进行的最好的一项采用了如下的方法，首先对三种主要算法的效率进行了标准化，然后在规定的效率水平对几种算法的检出功效进行比较（Ke et al. 2005）。这篇文章中表明，两种算法的应用功效之间存在显著性的差异，基于更为复杂的单体型的算法略优于连锁不平衡（LD）算法（Box 19-5），但是更重要的是，虽然此差异具有统计学显著性，但是其实际上还是十分接近的。也就是说，对于规定的 tag-SNP 的效率水平，两种算法的检出功效还是十分相似的。

**Box 19-5 连锁不平衡**

如果有两个 SNP，A 和 B，每个都有两个等位基因（分别为 A1、A2 和 B1、B2），有 4 种可能的单体型：A1B1、A1B2、A2B1、A2B2，这里我们写作 11、12、21 和 22。LD 统计描述的是两个 SNP 的单体型频率观察值和期望值之间的差异。基础统计量 $D$（Eq. 1）是指观察到的 A1B1 的单体型频率（$p_{11}$）与两个 SNP 基因型相互独立时的期望频率（$p_{A1} * p_{B1}$）的差值。

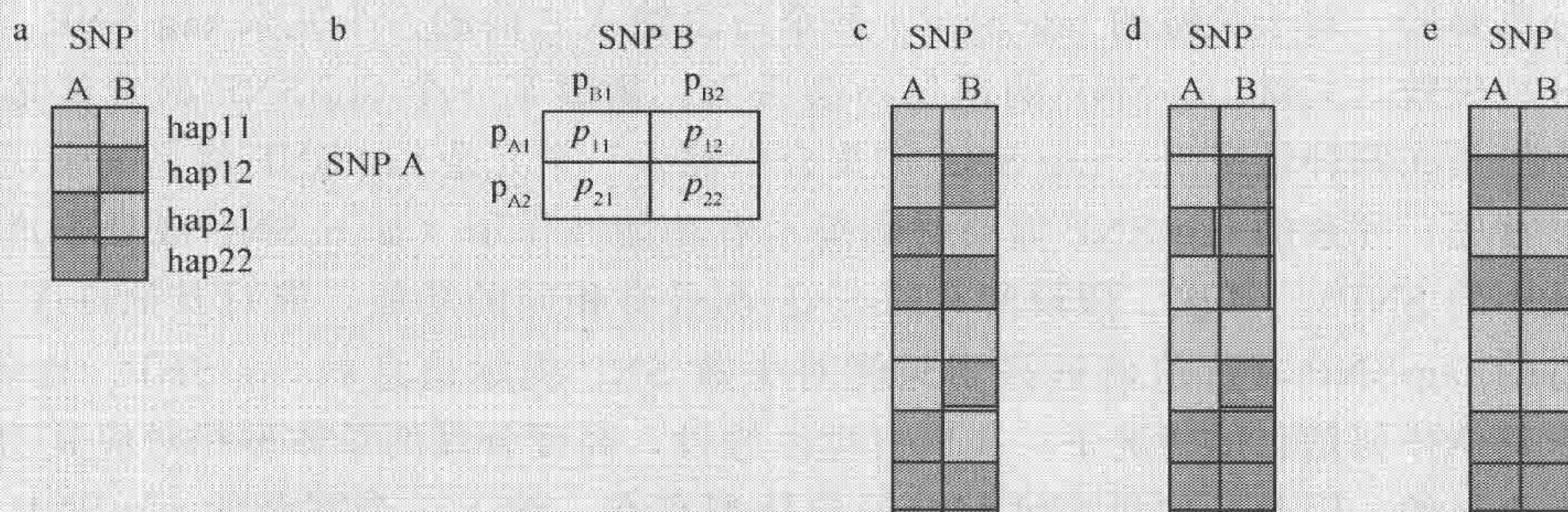
$$D = p_{11} - p_{A1} * p_{B1} = p_{11}p_{22} - p_{12}p_{21} \tag{1}$$

如果单体型频率恰巧和根据每个 SNP 的频率预测的结果一致，那么 $D=0$，这些 SNP 处于连锁平衡状态。对于一对 SNP，其次要等位基因频率（MAF）均为 50% 时，$D$ 的范围为 $-0.25 \sim 0.25$，这可能是 $D$ 最大的范围。但是，更低的 MAF 和（或）等位基因频率的失配将进一步限制此统计量的范围，这会使得比较成对 SNP 之间的 LD 值更难。为了对成对 SNP 间的 $D$ 值进行标准化，更常被使用的 LD 的统计量为 $D'$ 和 $r^2$（分别为 Eq. 2 和 3）。$D'$ 描述的是基于观察到的等位基因频率，观察到的 LD 相对于最大可能的背离，而 $r^2$ 描述的是两个 SNP 等位基因间的相关性。

$$D' = D/D_{\max} \tag{2}$$

$$r^2 = D^2/p_{A1}p_{A2}p_{B1}p_{B2} \tag{3}$$

当等位基因频率不一致时，甚至当 LD 尽可能的强的情况下，等位基因也不是完美相关。当完全 LD 的状态下，4 种可能的单体型中至少有一种不能被观察到，这与非重组区域一样，这时的 $D'=1$，并且 $r^2<1$。当等位基因频率一致时，基因型则可能完全相关，并且 $D'=1$，$r^2=1$，这反映了 A 和 B 两个多态位点的基因型完美相关。这个单体型中的等位基因将位于完美的连锁不平衡状态。




a. 用图表表示 4 种可能的单体型，黄色和蓝色分别标识两种等位基因。b. 观察到的四种单体型的频率可以用 2×2 的列联表来表示，其中 $p_{A1}$ 表示等位基因 A1 的观察频率，$p_{B1}$ 表示等位基因 B1 的观察频率，$p_{11}$ 是指 A1B1 单体型的观察频率。c. 此图表表示连锁平衡。d. 此图表表示“完全”连锁不平衡。这个例子中，没有出现单体型 21 [常见 A 等位基因（蓝色）]，罕见 B 等位基因（黄色）]。e. 此图表表示“完美”连锁不平衡。等位基因频率相同，两个 SNPA 和 B 的基因型呈现完美相关（见图版）

**从流行病学的观点看，tag-SNP 的选择的关键仅仅是某个遗传风险因子（可以是一个功能性 SNP 的基因型，或功能性 SNP 组成的单体型）和可测量的遗传学变异（一个 tag-SNP 的基因型或许多 tag-SNP 组成的单体型）之间的相关性。因此，大多数不依赖于单体型的 tag-SNP 选择算法使用的是 $r^2$ 或更复杂的 $r^2$ 演变的变量用于选择 tag-SNP，这样一个区域中常见变异的所有模式可以根据对一个 tag-SNP 进行检测或与某 tag-SNP 呈强相关的其他位点进行分析（或者一个 tag-SNP 组成的单体型）。**

如果在规定的效率水平，不同算法的功效近似，那么诸如实验设计和数据分析等实践问题的考虑则比理论问题的考虑要更为重要。实践问题的考虑重要可分为基因分型前和基因分型后两个部分。基因分型前主要的问题在于 tag-SNP 的灵活性。目前所有的算法在生物学上都是不完善的，其认为所有的 SNP 的功能都是等同的，仅对于所需要俘获已有的常见变异模式需要的 SNP 数目进行了优化。这很明显是具有缺陷的，因为可能一个 SNP 虽然具有很强的功能提示，但其却未被纳入最后的 tag-SNP。例如，ncSNP 在基因突变数据库中代表了不成比例的疾病-关联突变（Botstein and Risch 2003）。因此，如果某个 ncSNP 存在于基因组某区段中，由于其功能上可能存在的价值，其就应该被纳入作为一个 tag-SNP。因此，允许使用者强制纳入特定 tag-SNP 的算法显然是比单纯的传统算法具有更为优越。

第二种实践问题的考虑是 tag-SNP 的冗余。低效率的 tag-SNP 的应用将会增加一组 tag-SNP 中信息的冗余。这实际上也是一种优点，因为基于现有的技术平台，并非所有的 SNP 都能够相同效率的进行基因分型，占 10%～15%的 SNP 预期可能会分型失败。因此，如果 tag-SNP 中有一个基因分型失败，那么关注这一组 tag-SNP 的稳健性是十分重要的。对于关键的 tag-SNP，甚至可以纳入之前的一个冗余 tag-SNP，这样如果其分型失败，关键 tag-SNP 的信息不会丢失。衡量某一个 tag-SNP 的价值是十分复杂的，但是在两种情况下，tag-SNP 的集合中混入一个冗余 tag-SNP 将是一个明确的好策略。首先，冗余的 tag-SNP 需要明确提示有功能性作用（如 ncSNP 或处于进化高度保守域中的 SNP）。其次，假设所有的 SNP 都具有相同的功能，能对其他许多 SNP 提示信息的 tag-SNP 其价值就远高于只能对少数 SNP 提示信息的 tag-SNP。也就是说，当没有其他任何信息的情况下，我们有理由假设，高度分歧的单体型相对应其功能上也是高度分歧的。因此，如果高度保守的单体型存在，纳入一个冗余的 tag-SNP 对其进行识别是很可行的设计。由于相同原因，当使用基于连锁不平衡的 tag-SNP 方法时，俘获到大结构块中的 SNP 比小结构块中的 SNP 更为重要，因为大的结构块将能够标记出这些高度分歧的单体型。

关于 tag-SNP 的选择最后一个实际问题的考虑是关于数据的管理和记录的问题。高效率的算法常会使用 tag-SNP 组合的单体型对一些未分型的 SNP 进行基因型的推断。虽然，在基因分型中，这样的方式的优势在于更为经济，但是同时这样也对于使用者记录需要分析的 tag-SNP 的单体型带来了分析上的负担。纵然基因分型的费用正在快速减低，但是仍然应该对于 tag-SNP 进行彻底的优化，以达到对资源最为高效的实际

应用。

## 进入数据库和 tag-SNP 选择的应用

除理论上的考虑外，对于大多数研究人员，在应用 tag-SNP 算法时有两项要点：进入相对稠密的基因分型数据（或者更好的重测序后的数据）和进入 tag-SNP 选择的软件。获取基因型数据时有两个主要的数据库：HapMap 数据库（http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap/）HapMap 2003；Altshuler et al. 2005）以及 dbSNP 数据库（http://www.ncbi.nlm.nih.gov/SNP/）（Sherry et al. 1999）。HapMap 计划得到了三个地理分割人群的基因组详细信息，其包括欧洲人群、非洲人群和亚洲人群。因此，基于这些数据资源建立的工具对于目前基因组中未被进行重测序的区域的状态进行分析。然而，dbSNP 数据库是一个有多个项目来源的核心数据库（除了 HapMap 外，还包括其他靶向重测序项目），其应该能够最终取代 HapMap 作为核心数据库，从中获取的基因型数据可以用于 tag-SNP 的选择。

## 人群的特异性

无论采用何种算法，当在研究中挑选 tag-SNP 时，都需要注意所选用的人群和已进行基因分型的人群是否匹配。虽然人群间的变异分析提示常见 SNP 可能在所有主要人群中是具有多态性的，但是并非很难观察到在人群组间确实存在较大的等位基因频率的差异。核苷酸多态性以及 LD 的程度等参数在人群间也已经发现存在不同。总的来说，非洲人群中核苷酸多态性高于其他人群，这一点总是被误解为是因为非洲人群比其他人群历史更为久远而形成的。虽然，目前已经很明确现代人类在非洲居住的时间比其他大洲更长，但是非洲人群高度的序列多态性只能简单的解释为是在进化的时间线上，非洲人群的有效数目多于非洲外的人群。同样，非洲人群和其他人群相比连锁不平衡跨越的距离相对较短，这也与其具有较大的有效人群数目保持一致。

由于在人群之间存在等位基因频率和连锁不平衡的差异，至少用于进行 tag-SNP 选择的人群样本应该与研究人群样本在大洲的地理来源上保持匹配。在欧洲，等位基因频率的地理差异相对很小（Rosenberg et al. 2002），因此，HapMap 的欧洲样本可适用于大多数欧洲群体的 tag-SNP 的选择。HapMap 中的日本人群和中国人群之间，发现其等位基因频率的差异也相对较小，因此，这些样本中的数据信息可以适用于东亚人群中 tag-SNP 的选择。

实际情况下，我们往往并不是总能够匹配 tag-SNP 选择的来源人群和研究人群进行很好的匹配。例如，墨西哥-美国人群有明显的美国土著和欧洲祖先。如果在这样的一个研究人群中关注一些候选基因的话，可能只需要重测序研究人群中少数个体用于对 tag-SNP 的选择。如果重测序不可能实现，那么在每个祖居人群中可以独立进行 tag-SNP 的选择。例如，对于墨西哥-美国人群，可以采用 HapMap 亚洲和欧洲人群信息用

于 tag-SNP 的选择。这种情况下，事后进行等位基因频率的检查是很重要的：如果研究人群中某个 tag-SNP 观察到的等位基因频率不是分布在祖居人群的范围内，那么该区域可能会含有人群特异的变异，并且应该今后优先进行重测序。

HapMap 计划的第二阶段对三个主要的地理隔离人群中的数百万的 SNP 进行基因分型，包括欧洲人群、非洲西部人群（约鲁巴人人群）和东亚人群（中国和日本人群）。HapMap 计划的第三阶段将在别的人群中进行基因分型，但分型的位点仅为第二阶段中分型的一部分 SNP。因此，在这些人群中只有很少的变异图谱可以用于 tag-SNP 的选择。以与第三阶段人群匹配的研究人群为对象的研究人员应该记住这一点。在选择 tag-SNP 之前，研究人员应该考虑对于候选区域进行一些预先的重测序，在这些候选区域中匹配的第三阶段人群和第二阶段人群在等位基因频率上有显著的差异。

## 人群结构

tag-SNP 选择的目的是为了对一个人群中常见的变异模式进行有效的描述。但是，在某个祖居人群中的常见 SNP 可能是或可能不是别的祖居人群的常见 SNP。因而，对于混合人群进行 tag-SNP 的选择应该是一项挑战。一般来说，独立于祖居人群的数据资料或根据这些人群最佳的替代信息中进行 tag-SNP 的选择可以很好解决这个问题。例如，非洲-美国人平均约有 20%的欧洲血统，因此对于这个混合人群，应该根据欧洲和非洲的 HapMap 人群数据进行 tag-SNP 的选择。西班牙人在这个问题上则更为复杂，对于西班牙人群的研究可能需要在非洲、欧洲和亚洲人群中进行 tag-SNP 的选择，特别是对加勒比海的西班牙人群更是这样。最近已经有了一些具体的策略以帮助对多来源的混合人群进行一系列 tag-SNP 的选择（Howie et al. 2006）。

## 选择 tag-SNP：使用者手册

当我们选择一系列 tag-SNP 时，需要确定一个目标靶区域（可以是一个基因，也可以是整个基因组），在一个适当的 tag-SNP 选择人群的初步实验得到基因型的数据，以及基于这些信息进行 tag-SNP 选择的工具。尽管一些使用者会从研究人群中挑选一个代表性的样本群体进行重测序，从而得到他们自己初步的基因型结果，但是大多数使用者更倾向于希望使用已有的初步实验的信息来选择适当的 tag-SNP。dbSNP 数据库是关于基因型信息的标准数据库，但是直接从 dbSNP 数据库中提取原始的基因型信息并不是很简单的事情。因此，对于使用者而言能够很简单的进行基因型数据提取的工具是十分重要的，并且这种工具要能够方便不同格式的数据相互转换。

### 使用 Genome Variation Server

从 dbSNP 数据库中提取基因型信息的工具之一是 Genome Variation Server(GVS, http://gvs.gs.washington.edu/GVS/)。这个工具可以使得使用者使用多种参考方式

提交至基因组，以进行查询：使用基因组中的位标，基因的名称，人类基因组组织（HUGO）的基因名称，或者一个锚定 SNP（图 19-3）。当面对一个基因簇时，基因组的位标是非常有用的；而对于选择兴趣区域时，使用锚定 SNP 是十分有用的策略；但是一般来说，大多数查询还将会是从一个已想好的候选基因开始。一旦数据提取出来后，GVS 也将提供内置的基于 LD 选择而考虑单体型的 tag-SNP 选择算法（Carlson et al. 2004）。



图 19-3　Genome Variation Server 的主页。此程序可用于通过不同的参数的设置搜索感兴趣的基因

假设我们对瘦素基因 Leptin 感兴趣（HUGO 基因名称为 LEP）。如果点击“gene id”字段名，将提示输入基因的 id（图 19-4）。所获取的序列的默认范围为该基因已知最长的 mRNA 的 refseq 序列，但是您也可以对特定的序列上游（转录起始位点的 5′端）或下游（poly（A）信号的 3′端）范围进行限制。在我们这个例子中，我们将在基因名称中输入 LEP，上游 5000bp 和下游 2500bp，然后点击“搜索”。

提交的查询会返回一个人群列表，在这些人群中已经在 dbSNP 数据库中有该特定区域的基因型数据的报道（图 19-5）。您可以通过点击人群的名称进行链接，这样可以对每个人群的详细情况进行更加深入的了解。我们这个例子中，将选择 PGA _

GVS: Genome Variation Server

Sponsored by SeattleSNPs

About GVS
How to Use GVS
Build Notes
Contact Us

Gene Name Search

gene name: LEP

Beyond Your Target (optional)

upstream (# of bases): 5000

downstream (# of bases): 2500

search

reset

Tuesday, April 25, 2006

图 19-4　使用 GVS 系统的实例，输入数据以搜索 Leptin 基因的基因型数据。搜索也可以设置对于编码区外侧的序列进行

YORUB-PANEL，它是由 SeattleSNPs PGA 在这个区域中对 HapMap 的约鲁巴人进行重测序后的结果。选择合适的人群后，在屏幕的页面下部可以进行参数的设置，可以对数据显示和/或采用 LDselect 对 tag-SNP 的选择进行调整（图 19-6）（Carlson et al. 2004）。

选择 tag-SNP 前，将“Display SNPs By”设置为“Text/Image”，并且点击“Display Genotypes”。这时会弹出两个窗口，一个显示的是数据文本格式的 prettybase 文件，另一个显示的是图形化可视的基因型（图 19-7）。Prettybase 文件是导入 LDselect 和 multipop-TagSelect 的原始数据文件，其中的数据排成四列，之间用 Tab 制表符隔开：多态性位点，个体 ID，等位基因 1，等位基因 2。将 prettybase 存为文本文件今后备用。另外，如果你有自己分型的基因型结果希望上传至 GVS 的话，你将需要将你的数据转变为这种格式。将 prettybase 格式转换为 ped 格式（HaploView）时，转换的工具可以在下列地址进行下载，http://theta.ncifcrf.gov/gbrowse/start_db/prettybase_to_linkageformat.zip，将 ped 格式转换为 prettybase 格式（HaploView）时，转换的工具可以在下列地址进行下载。http://www.pharmgat.org/pharmgat.org/Documentation/help/pedtopb。

再回到 tag-SNP 的选择，退回至 GVS tag-SNP 参数设置的页面（图 19-6）。这里你可以在 Leptin 基因数据中对于基于 LDselect 的 tag-SNP 选择的相关参数进行设置。“Output SNPs By”可以被设置为当前基因组序列中的“SNP_position”，或 dbSNP 数

GVS: Genome Variation Server

Sponsored by SeattleSNPs

Gene Name: LEP map expand search

Gene ID: 3952

Chromosome 7: 127475281 - 127491632 (+)

Total chromosome span: 127470281 - 127494132

**Select Population(s)**

Check at least one population. You may select up to 4 populations if all those selected have individuals in common.

| Select | Number Variations | Number Genotypes | Population | Submitter | Check to Select Only Unrelated Individuals |
|---|---|---|---|---|---|
| ☑ | 90 | 2160 | PGA_YORUB-PANEL | PGA-UW-FHCRC | |
| ☐ | 90 | 2070 | PGA_CEPH-PANEL | PGA-UW-FHCRC | |
| ☐ | 53 | 3240 | HapMap-JPT | CSHL-HAPMAP | |
| ☐ | 53 | 3240 | HapMap-HCB | CSHL-HAPMAP | |
| ☐ | 52 | 4680 | HapMap-YRI | CSHL-HAPMAP | ☐ |
| ☐ | 50 | 4500 | HapMap-CEU | CSHL-HAPMAP | ☐ |

图 19-5 根据图 19-4 中输入的搜索条件，GVS 系统生成的人群选择界面

**Set up parameters for display and analysis**

| Data Output and Display | | | |
|---|---|---|---|
| Output SNPs By | SNP_Position | Display SNPs By | Table/Image |
| Filtering SNPs | | | |
| Allele Frequency Cutoff (%) | 0 | No Monomorphic Sites | ☐ |
| Clustering in Graphic Display | | | |
| Cluster SNPs | ☐ | Cluster Samples | ☐ |
| Selecting Tag SNPs | | | |
| $r^2$ Threshold (0.0-1.0) | 0.80 | Data Coverage (%) for TagSNPs | 85 |
| | | Data Coverage (%) for Clustering | 70 |
| Color-Coding For LD Plot | | | |
| LD Minimum (0.0-1.0) | 0.1 | LD Maximum (0.0-1.0) | 1.0 |

**Display Results**

display genotypes → display tag snps → display image linkage disequilibrium → display snp summary →

图 19-6 设置用于 tag-SNP 选择的参数。GVS 系统提供了大量的选项，研究人员可以对搜索的参数选项和数据显示进行设置

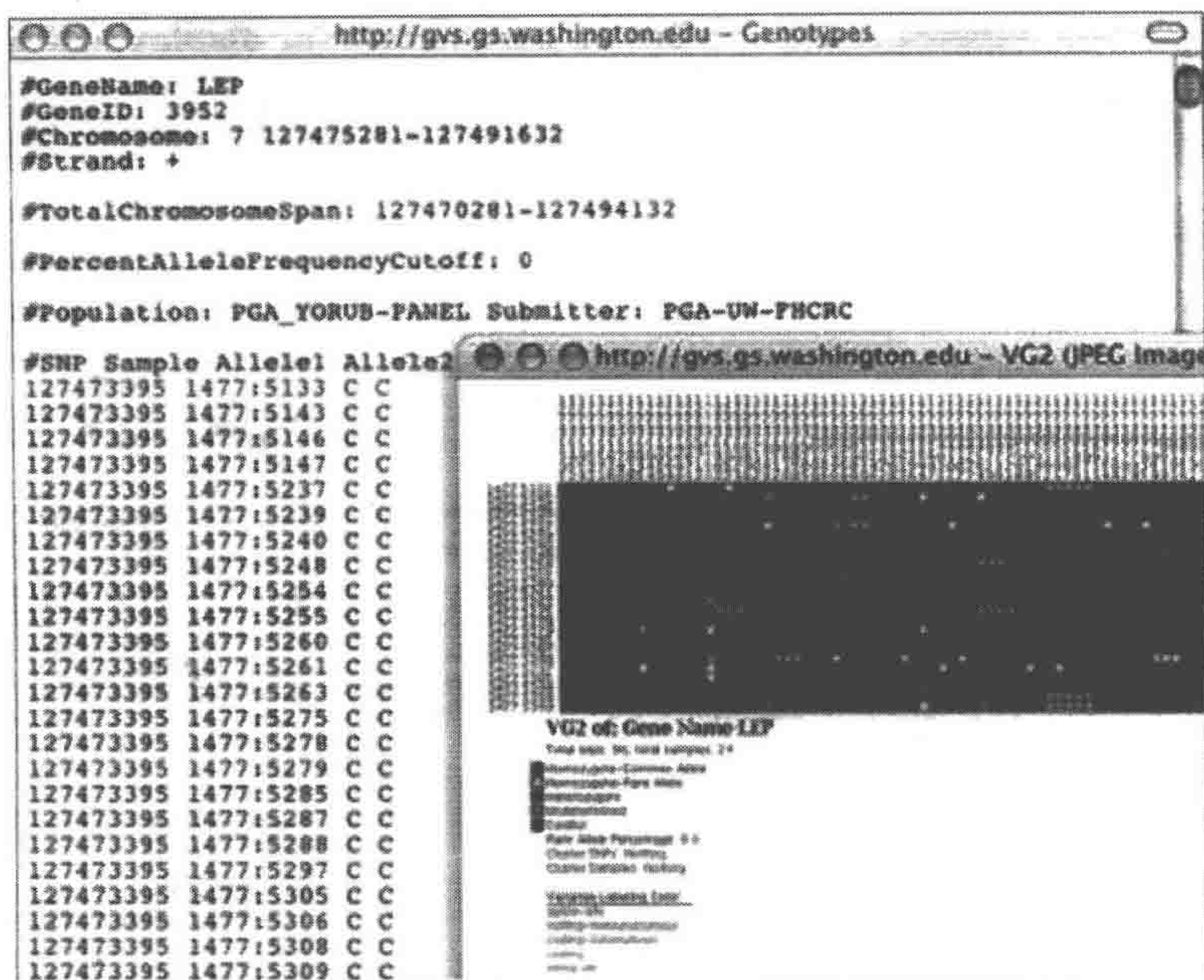



图 19-7　在 GVS 中可视化的基因型。除了图形化的显示外，GVS 的弹出窗口还以文本格式（prettybase 格式）显示基因型的数据

据库中各个多态位点的参考 SNP 编号“RS _ ID”。不同研究之间使用 RefSNP 代码则可以更容易的实现同一位点的相互对应，在任何关联研究的报道中都常常被使用，但是，使用每个碱基的坐标更容易理解基因排列的相对位置。这时，将这个选项设置为“SNP _ Position”。下一个需要设置的参数是“Allele Frequency Cutoff (%)”。对于可能的样本量的情况，进行检验效能的估计，再根据其对研究中的最小等位基因频率进行选择设置，但是一般来说，应该设置为 5%或者 10%。这里设置为 5%。复选框“No Monomorphic Sites”是让使用者排除任何非多态的 SNP（通常一个 SNP 可能在另外别的一个群体中是有多态性的）被后续进行显示/分析。最后，tag-SNP 的选择参数需要进行设置：$r^2$ 的阈值也应该根据具体样本量的情况估计的检验效能进行确定。默认值为 0.8 对于候选基因分析具有可接受的严格程度，一般来说，不应该设置此参数低于 0.5。我们等下将回到“Data coverage for Tag-SNPs”。现在，先点击一下页面底部的“display tag-SNPs”按钮。结果在图 19-8 中列出。

tag-SNP 选择的结果在两个框架中显示：第一个（图 19-8）框架中显示的是一个新的 prettybase 文件，其中的 SNP 被进行了重排，具有强连锁不平衡的 SNP 邻近排列。这张图顶部的黑横条表示 SNP 存在的箱块，星号表示每个块中信息量等价的 tag-SNP。每个箱块中，只需要对一个星号标识的 SNP 进行基因分型。SNP 用彩色进行标记的方法在图例中已注明。例如，第三个箱块中有一个非同义 SNP（氨基酸发生改变），用红色表示。当在每个箱体中选择 tag-SNP 时，由于这些红色标记的 SNP 可能具有功能提示，因此可以得到更加的关注：由于箱块 3 中的非同义 SNP 可能在功能提示上比非编

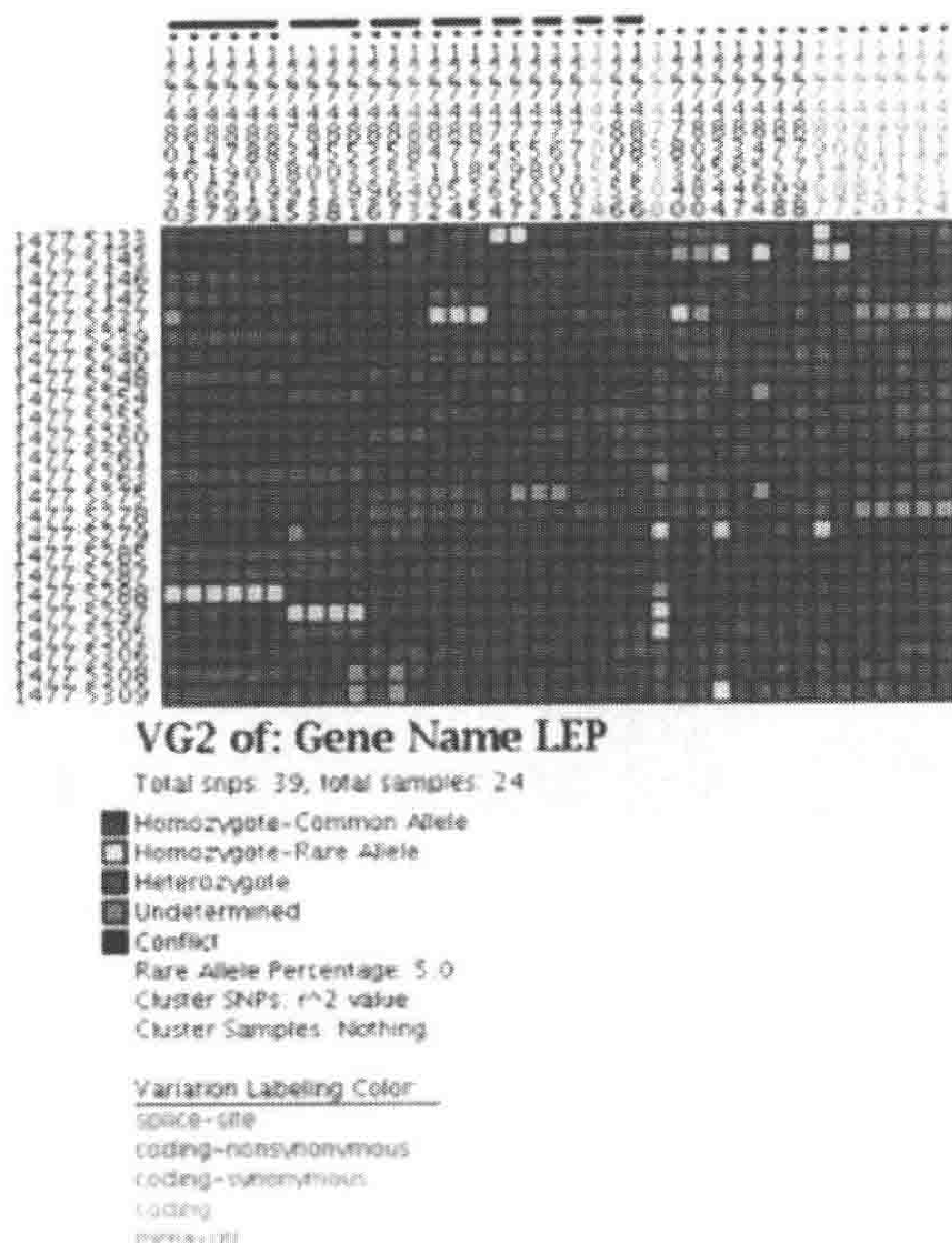



图 19-8　在约鲁巴的样本中挑选 Leptin 基因的 tag-SNP。这些结果是对应于图 19-5 的选择界面中特定设置的参数而生成的（细节信息请见原文中）。这些 SNP 被进行了重新排列，处于强的 LD 关系的相互临近。这张图的顶部，黑色的横条表示 SNP 存在的箱块，星号表示每个块中信息量等价的 tag-SNP。SNP 用彩色进行标记的方法在图例中已注明

码 SNP 具有更多的意义，因此，在进行基因分型前其应该被优先进行选择。

回到“coverage”的选项，“Data Coverage for tag-SNPs”可以用来定义对于一个 tag-SNP 来说，其基因型缺失的数据可以允许达到什么样的程度。这一点是很重要的，因为有缺失数据的 SNPs 有时候更容易被选作 tag-SNP（如第二个箱块中左起的 tag-SNP127485391）。将此参数设置为 100 可以提前排除选择此 SNP 作为 tag-SNP，并且能够使得第二个箱块中的结构存在更多的选择。“Data Coverage for clustering”的选项较为类似，其实际上将信息过多缺失的 SNPs 设置为其独立的箱块而存在。

## Haploview 和 Tagger

另外一个提取基因型数据并且进行 tag-SNP 选择的工具是 Haploview（http://www.broad.mit.edu/mpg/haploview/download.php；也可参见第 21 章）。Haploview 所能完成的分析功能绝大部分和 GVS 相似，只进行了很小的改动。首先，Haploview 是一个独立的 Java 小应用程序，其操作的数据需要先整理成 HapMap 的格式。因此，为了使用 Haploview 进行 tag-SNP 的选择，必需首先从 HapMap 的网站上下载相应区

域的基因型数据（图 19-9）。在我们这个例子中，我们将使用相同的片段：在 Landmark or Region 的框中，我们输入“Chr7：127475281.. 127491632”，然后点击查找。你也可以使用 HUGO 基因名称来搜索具体的靶区域。搜索结果在图 19-10 中列出。现在，在“Report & Analysis”下拉菜单中，选择“Download SNP Genotype Data”，然后点击“Configure”。选择适当的人群（如 YRI 表示约鲁巴人）后，选择“Save to Disk”，然后点击“Go”。文件将以“dumped _ region”文件名进行存储，你需要对其存储位置进行设置。

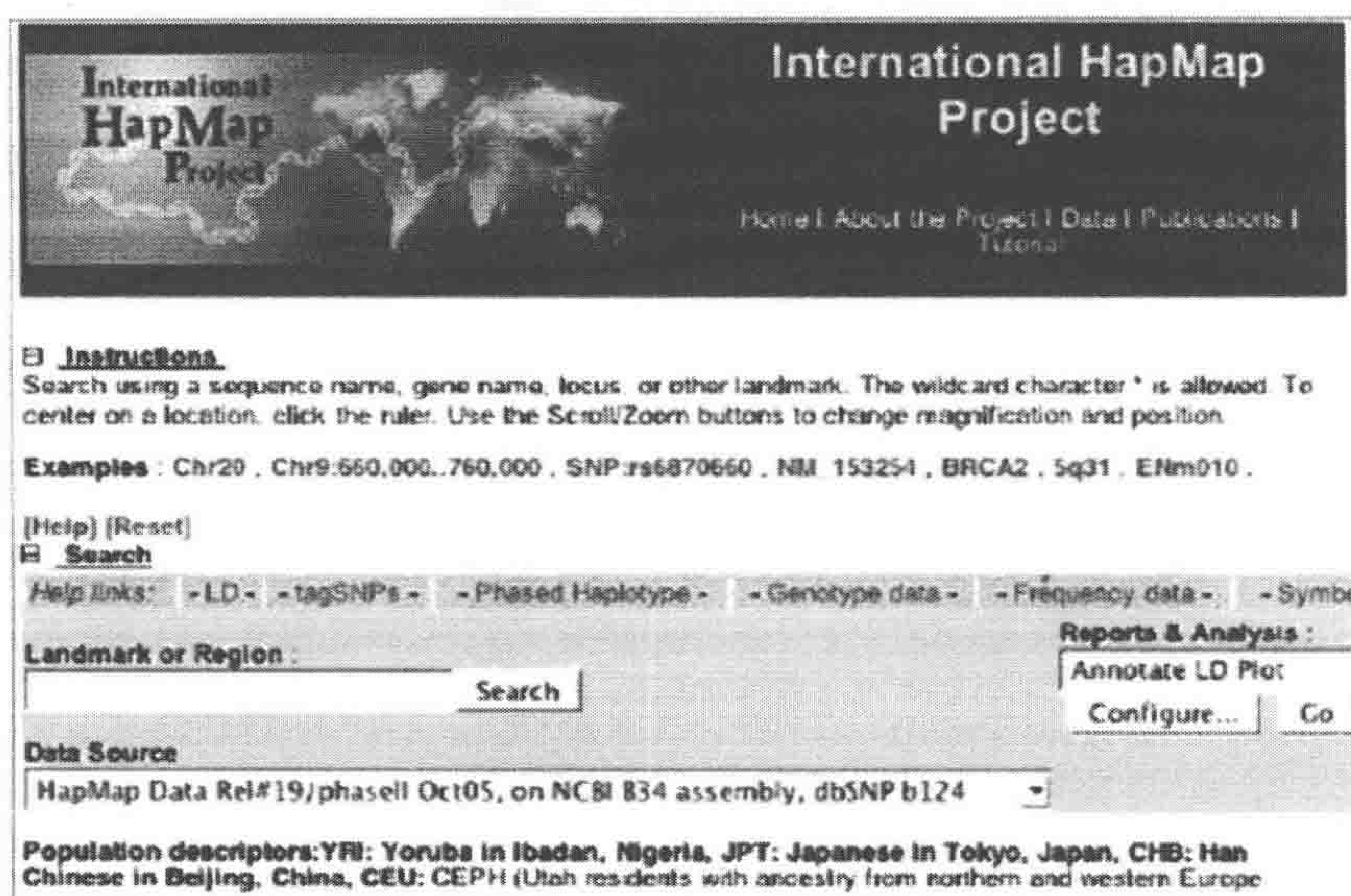



图 19-9　HapMap 计划网站可以用来下载基因型的数据信息

现在打开 HaploView。默认屏幕界面将有三个选项：选择“Load Hapmap Data”。如果看不见这些选项，选择文件-打开基因型数据，然后找到你的文件。一旦文件被打开（图 19-11），选择“Tagger”，然后你将可以进行 Tag SNP 的选择。“pairwise taggin only”的选项对于 LDselect 的算法非常重要（Calson et al. 2004），并且，“Force Include”和“Force Exclude”两个选项常用于对将要进行基因分型（如功能性 SNP）或不能进行分型的 SNP（如 SNP 位于序列重复元件中，大多数高通量基因分型平台都不适用）进行定义。“aggressive tagging”选项，可以允许未分型的 SNP 根据其他 SNP 组成的单体型进行基因型的推测（de Bakker et al. 2005）。这能够减少所需要的 tag-SNP 的数目，但是其缺点在于任何分型失败的 tag-SNP 将需要更高的花费。选择“aggressive tagging：Use 2-marker haplotype”，然后点击“Run Tagger”。结果在图 19-12 中给出。

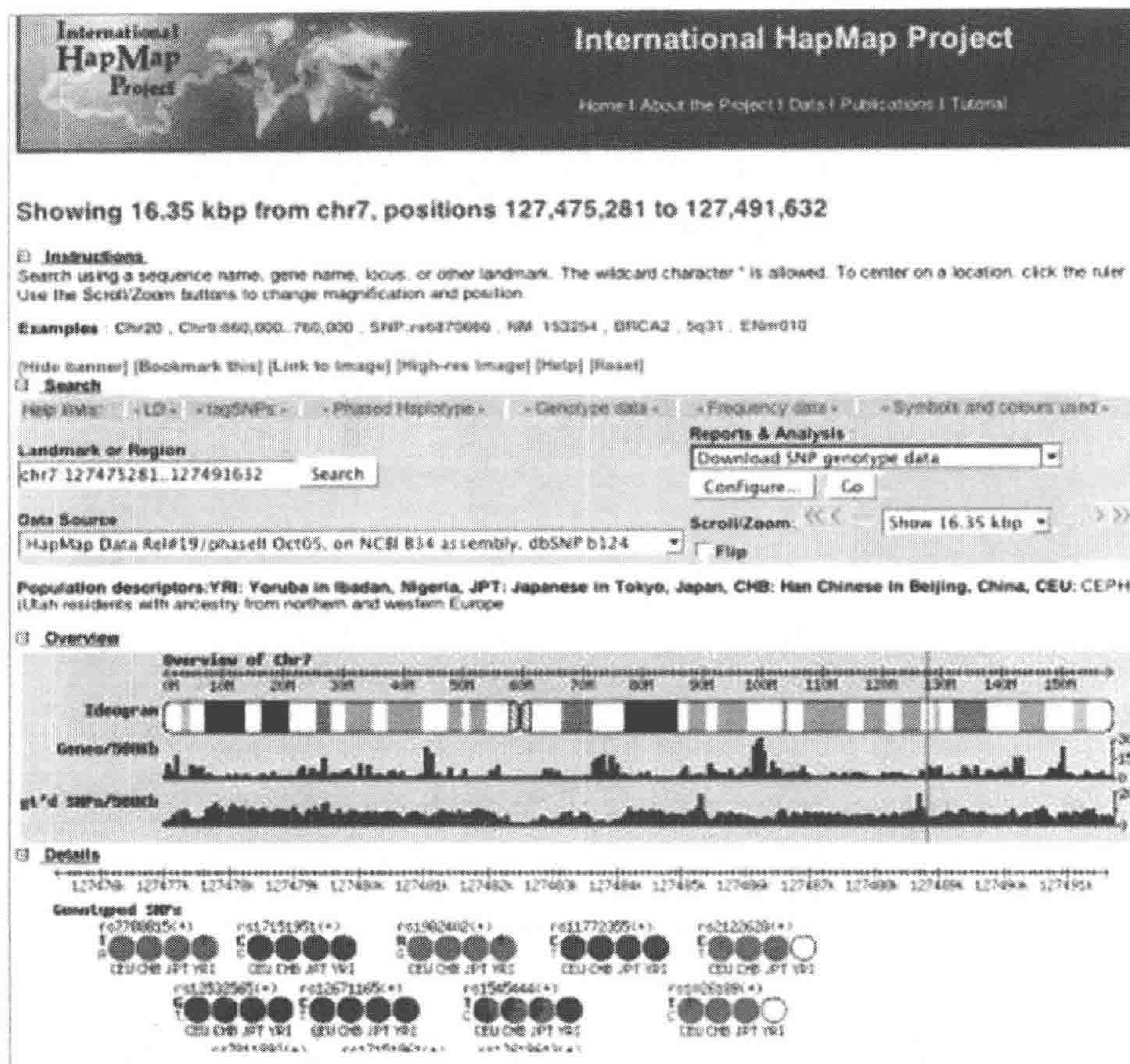



图 19-10 从 LEP 基因区域中获取 HapMap 的基因型数据的实例

## 比较 GVS 和 Haploview

Haploview 和 GVS 这两个工具分别有各自的优势和缺点。二者都基于基本的配对 $r^2$ 进行 tag-SNP 的选择，并且二者都可以对连锁不平衡的模式（这一点在这里没有进行探讨）进行可视化的分析。GVS 工具最显著的优点是其可以使用来自 dbSNP 数据库的基因型数据进行处理。虽然 HapMap 对于大多数基因组是最好的选择，但是对于 tag-SNP 选择的金标准是重测序的数据结果，其来自于 SeattleSNPs，HapMap ENCODE，先天免疫基因组应用项目以及今后可能的其他研究组，对于大量数目的候选基因进行重测序的结果。使用 GVS 工具另一个显著的优点，在于根据最初功能预测的类型分类，SNP 可使用多种颜色进行标识。虽然 Haploview 工具强制的“include”和“exclude”功能也是十分有用的，Haploview 工具最显著的优点是其具有的强化某 SNP 为标签 SNP 的特点。或许，在未来的发展中，两个工具都将继续完善，并且整合入新的特点和用途。

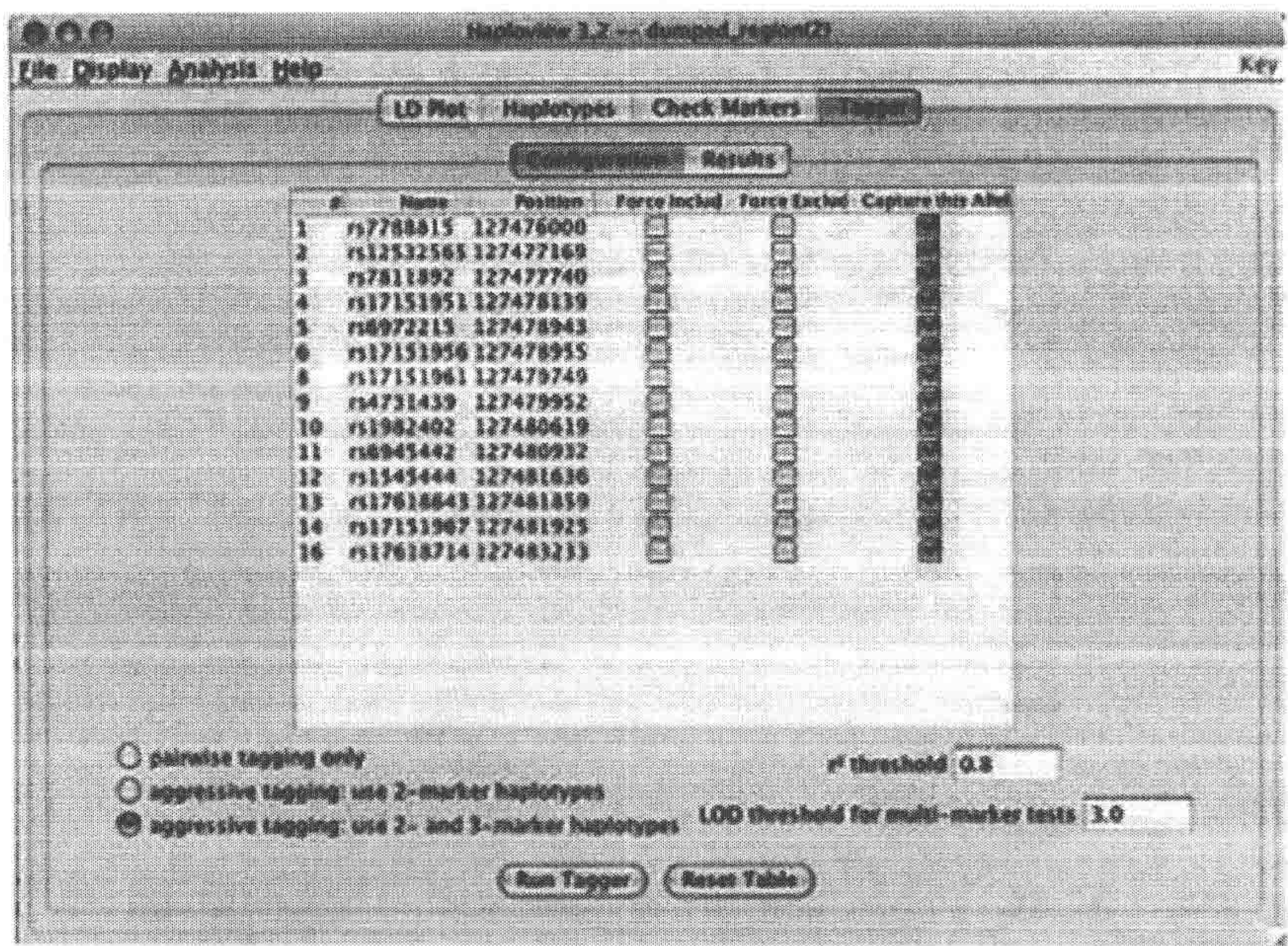



图 19-11　Haploview 程序可以操作从 HapMap 数据库中调取的基因型的信息。“Tagger”功能可以对基因型进行分析，以进行 tag-SNP 的选择

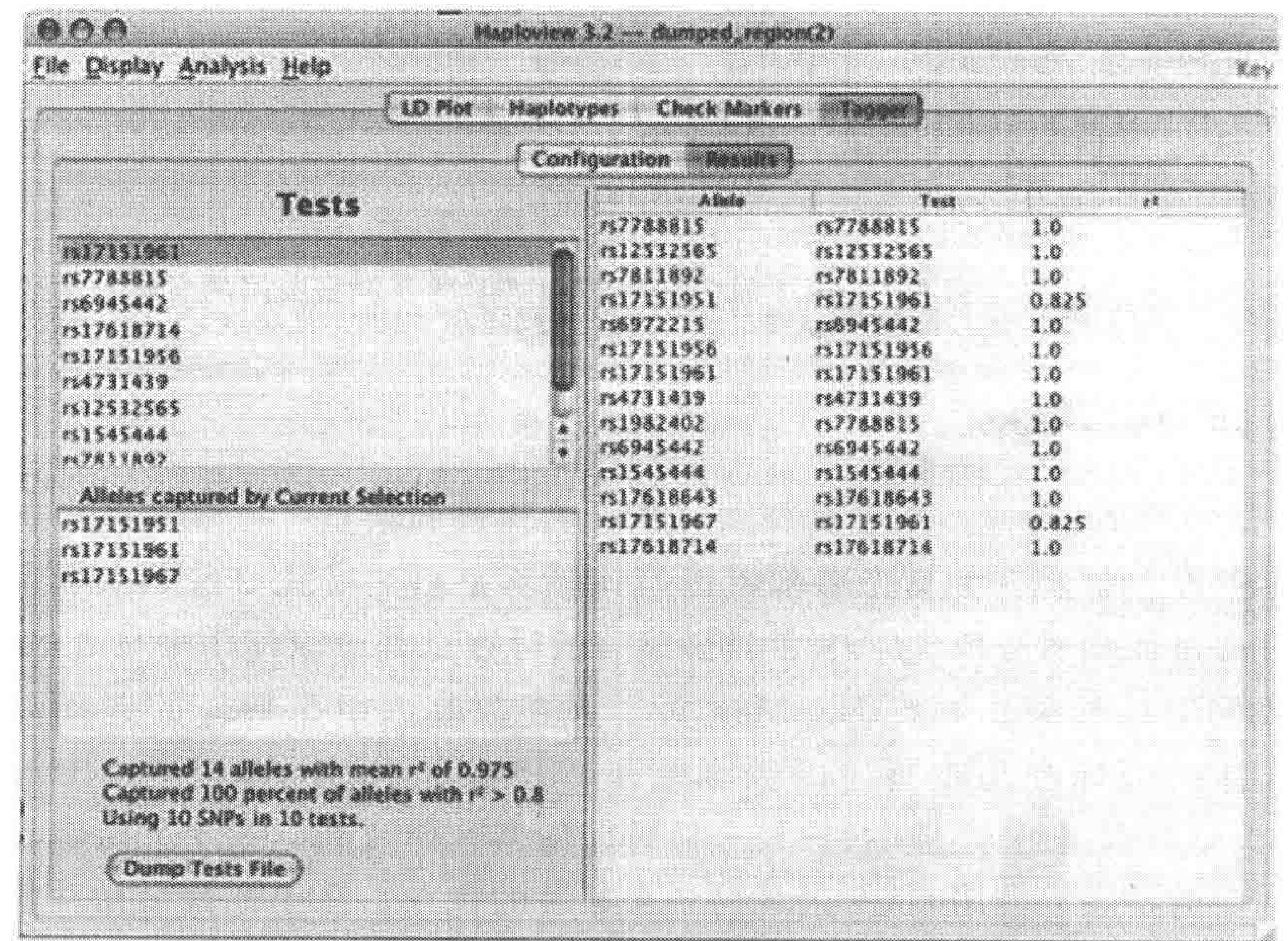



图 19-12　Tagger 输出约鲁巴 HapMap 样本 Leptin 基因区域的 tag-SNP 选择结果。在此输出结果中，一组 tag-SNP 作为“Tests”列在左侧，右边面板中列出了此区域中的 SNP（“Allele”列）和每个位点被 tag-SNP 俘获的情况（“Test”列）

## tag-SNP 的未来

现在从技术上可以实现对于一个样本进行几十万个 SNP（甚至百万个）的基因分型（Matsuzaki et al. 2004；Steemers et al. 2006）；但是，当面对样本量很大时，这种策略在价格上过于昂贵。因此，在减少 SNP 的数目同时也要将丢失的信息降到最低点成为了候选基因分析时的一项重要挑战。这从根本上是一个经济问题，并且基因型是非常灵活的，认识到这一点非常重要。为了对一个基因、一条通路甚至全基因组进行关联分析，研究者的预算主要取决于每个样本的花费，而不是每个基因型的花费。也就是说，tag-SNP 的选择主要是为了减少研究中需要进行基因分型的数目，使得开销满足最初的预算。最后，可能对全基因组进行重测序的花费将接近全基因组 tag-SNP 进行基因分型的花费，到那时，再对 tag-SNP 进行挑选则变得过时了。

## 参考文献


Altshuler D., Brooks L.D., Chakravarti A., Collins F.S., Daly M.J., and Donnelly P. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Bhangale T.R., Rieder M.J., Livingston R.J., and Nickerson D.A. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14:** 59–69.

Botstein D. and Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 Suppl:** 228–237.

Cardon L.R and Abecasis G.R. 2003. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19:** 135–140.

Carlson C.S., Eberle M.A., Rieder M.J., Yi Q., Kruglyak L., and Nickerson D.A. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74:** 106–120.

Collins F.S., Guyer M.S., and Chakravarti A. 1997. Variations on a theme: Cataloging human DNA sequence variation. *Science* **278:** 1580–1581.

de Bakker P.I., Yelensky R., Pe'er I., Gabriel S.B., Daly M.J., and Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* **37:** 1217–1223.

Ding K., Zhou K., Zhang J., Knight J., Zhang X., and Shen Y. 2005. The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. *Mol. Biol. Evol.* **22:** 148–159.

Gabriel S.B., Schaffner S.F, Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., et al. 2002. The structure of haplotype blocks in the human genome. *Science* **296:** 2225–2229.

Halldorsson B.V., Istrail S., and De La Vega F.M. 2004a. Optimal selection of SNP markers for disease association studies. *Hum. Hered.* **58:** 190–202.

Halldorsson B.V., Bafna V., Lippert R., Schwartz R., De La Vega F.M., Clark A.G., and Istrail S. 2004b. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res.* **14:** 1633–1640.

Halperin E., Kimmel G., and Shamir R. 2005. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics* (Suppl. 1) **21:** i195–i203.

Howie B.N., Carlson C.S., Rieder M.J., and Nickerson D.A. 2006. Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Hum. Genet.* **120:** 58–68.

International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.

Jeffreys A.J., Kauppi L., and Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29:** 217–222.

Johnson G.C., Esposito L., Barratt B.J., Smith A.N., Heward J., Di Genova G., Ueda H., Cordell H.J., Eaves I.A., Dudbridge F., et al. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29:** 233–237.

Ke X. and Cardon L.R. 2003. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19:** 287–288.

Ke X., Durrant C., Morris A.P., Hunt S., Bentley D.R., Deloukas P., and Cardon L.R. 2004. Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* **13:** 2557–2565.

Ke X., Miretti M.M., Broxholme J., Hunt S., Beck S., Bentley D.R., Deloukas P., and Cardon L.R. 2005. A comparison of tagging methods and their tagging space. *Hum. Mol. Genet.* **14:** 2757–2767.

Kimura M. and Ota T. 1973. The age of a neutral mutant persisting in a finite population. *Genetics* **75:** 199–212.

Kruglyak L. and Nickerson D.A. 2001. Variation is the spice of life. *Nat. Genet.* **27:** 234–236.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., et al. (International Human Genome Sequencing Consortium). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lin Z. and Altman R.B. 2004. Finding haplotype tagging SNPs by use of principal components analysis. *Am. J. Hum. Genet.* **75:** 850–861.

Matsuzaki H., Dong S., Loi H., Di X., Liu G., Hubbell F., Law J., Berntsen T., Chadha M., Hui H., et al. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods.* **1:** 109–111.

McVean G.A., Myers S.R., Hunt S., Deloukas P., Bentley D.R., and Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581–584.

Meng Z., Zaykin D.V., Xu C.F., Wagner M., and Ehm M.G. 2003. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.* **73:** 115–130.

Ng P.C. and Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res.* **11:** 863–874.

———. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31:** 3812–3814.

Patil N., Berno A.J., Hinds D.A., Barrett W.A., Doshi J.M., Hacker C.R., Kautzer C.R., Lee D.H., Marjoribanks C., McDonough D.P., et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294:** 1719–1723.

Risch N. and Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* **273:** 1516–1517.

Rosenberg N.A., Pritchard J.K., Weber J.L., Cann H.M., Kidd K.K., Zhivotovsky L.A., and Feldman M.W. 2002. Genetic structure of human populations. *Science* **298:** 2381–2385.

Sebastiani P., Lazarus R., Weiss S.T., Kunkel L.M., Kohane I.S., and Ramoni M.F. 2003. Minimal haplotype tagging. *Proc. Natl. Acad. Sci.* **100:** 9900–9905.

Schulze T.G., Zhang K., Chen Y.S., Akula N., Sun F., and McMahon F.J. 2004. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum. Mol. Genet.* **13:** 335–342.

Sherry S.T., Ward M., and Sirotkin K. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation [In Process Citation]. *Genome Res.* **9:** 677–679.

Steemers F.J., Chang W., Lee G., Barker D.L., Shen R., and Gunderson K.L. 2006. Whole-genome genotyping with the single-base extension assay. *Nat. Methods.* **3:** 31–33.

Stram D.O., Haiman C.A., Hirschhorn J.N., Altshuler D., Kolonel L.N., Henderson B.E., and Pike M.C. 2003. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55:** 27–36.

Templeton A.R., Clark A.G., Weiss K.M., Nickerson D.A., Boerwinkle E., and Sing C.F. 2000. Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* **66:** 69–83.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wall J.D. and Pritchard J.K. 2003a. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* **73:** 502–515.

———. 2003b. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **4:** 587–597.

Watterson G.A. 1976. Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor. Popul. Biol.* **10:** 239–253.

Weale M.E., Depondt C., Macdonald S.J., Smith A., Lai P.S., Shorvon S.D., Wood N.W., and Goldstein D.B. 2003. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene *SCN1A*: Implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73:** 551–565.

Zhang K. and Jin L. 2003. HaploBlockFinder: haplotype block analyses. *Bioinformatics* **19:** 1300–1301.

Zhang K., Qin Z., Chen T., Liu J.S., Waterman M.S., and Sun F. 2005. HapBlock: Haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* **21:** 131–134.

Zhang K., Qin Z.S., Liu J.S., Chen T., Waterman M.S., and Sun F. 2004. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res.* **14:** 908–916.

# 互联网资源

**http://gvs.gs.washington.edu/GVS/** The Genome Variation Server is a useful tool to extract genotypes from the dbSNP using multiple frames of reference. It also offers built-in implementation of the LDselect haplotype-free tag-SNP selection algorithm.

**http://theta.ncifcrf.gov/gbrowse/start_db/prettybase_to_linkageformat.zip** A converter from prettybase to ped format (HaploView).

**http://www.broad.mit.edu/mpg/haploview/download.php** Haploview is a useful tool for extracting genotype data and running tag-SNP selection. Unlike GVS, Haploview is a stand-alone Java Applet that manipulates genotype data in HapMap format.

**http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap/** The HapMap project is a major repository for genotype data providing remarkably detailed data sets in three geographically discrete human subpopulations: Europeans, Africans, and Asians. Tools built upon this resource are currently the state of the art for those regions of the genome that have not been targeted for resequencing.

**http://www.ncbi.nlm.nih.gov/SNP/dbSNP** is the central standard repository for genotyping information, including HapMap as well as other targeted resequencing projects.

**http://www.pharmgat.org/pharmgat.org/Documentation/help/pedtopb** A converter from ped format to prettybase format.

# 20 使用 Tagger 和 HapMap 软件挑选和评价 tag-SNP

Paul I.W. de Bakker

*Program in Population and Medical Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142*



## 简介

国际人类基因组单体型图计划已经构建了一个开放的数据库，其包含超过 300 万个已分型的 SNP，这些 SNP 覆盖了 4 个人群的 270 个基因组（International HapMap Consortium 2005）。通过提供 SNP 的基因组物理位置和相互的连锁关系，HapMap 有助于设计全基因组关联研究和数据分析。本章介绍怎样使用 Tagger（http://www.broad.mit.edu/mpg/tagger/；de Bakker et al. 2005）在线软件从 HapMap 数据库中挑选和评价 tag-SNP。此外，Tagger 已经在 Haploview 软件里使用（http://www.broad.mit.edu/mpg/haploview/；Barrett et al. 2005），这部分内容将在第 21 章里介绍。在我们介绍使用 Tagger 前，首先阐述使用不完备数据如 HapMap 的三个重要问题和挑选 tag-SNP 的常用方法。

### HapMap 数据库如何提供常见变异及那些没有被收录的变异的信息

HapMap-ENCODE 数据中的常见变异信息基本被确定。HapMap 计划的第二阶完成后，它收录了 CEU（犹他州的北欧和西欧后裔）和 CHB＋JPT（中国北京汉族和日本东京人群）的常见变异，其中等位基因频率≥5％的 SNP 达 94％，尼日利亚伊巴丹约鲁巴人 81％常见 SNP 满足 $r^2 \geq 0.8$。虽然非洲后裔人群的少部分变异没有被收录，但 HapMap 已经包括了大部分常见 SNP 的信息（International HapMap Consortium 2005）。并不令人吃惊，HapMap 并未把非常见变异都收录进去（如稀有突变）。

**当从 HapMap 数据库中仅挑选一组 SNP（tag-SNP）时，怎样减少使用的常见变异**

邻近的遗传标记在染色体上广泛存在着结构关系。通过开发 SNP 间的配对关系能获得巨大的效率。近期我们证明了一种基于单体型标签的方法（如 aggressive），可以在不减少检验效能情况下转换这种平衡（de Bakker et al. 2005）。用检验效能来评价关联似乎与参考数据库的完整情况存在明显关系。

**从 HapMap 数据库中挑选出的 tag-SNP 在其他群体里捕获常见变异的能力如何**

大量研究已经评价了 tag-SNP 在不同群体中的转移能力。(Nejentsev et al. 2004; Ahmadi et al. 2005; Mueller et al. 2005; Ramirez-Soriano et al. 2005; de Bakker et al. 2006a, b; Gonzalez-Neira et al. 2006; Huang et al. 2006; Montpetit et al. 2006; Ribas et al. 2006; Stankovich et al. 2006)。总体来讲，这些研究显示从 Centre d'Etude du Polymorphisme Humain collection 数据库里（如 HapMap CEU 人群）挑出的 tag-SNP 在其他欧裔人群中同样能有效地捕获常见变异。从 HapMap CHB 和 JPT 挑出的 tag-SNP 在其他亚洲来源群体中也能获得相似的结果。当然能否在人群里有效传递，取决于该群体染色体局部连锁不平衡（LD）的程度和近来人群的混杂程度（如非洲人和美国人的混杂）。鼓舞人心的是 HapMap 数据在全球多个群体都能捕获常见变异（Conrad et al. 2006; de Bakker et al. 2006b）。

## tag-SNP 的挑选

tag-SNP 挑选的第一步是定义一组等位基因，它们将被 tag-SNP 捕获并检测基因型和表型的关系。定义等位基因可以通过在 HapMap 中输入具体的染色体物理位置可以完成（如染色体 6：32000000～33000000 来定义 MHC 区段的等位基因。目前，tagger 在线服务器采用 HapMap 第 21 次开放的镜像数据，数据物理位置参照 NCBI 颁布的第 35 版基因组信息（UCSC 版 hg17），但是随着新的 HapMap 数据公布，物理位置可能被更新。这一功能对于一个包含多个候选基因的研究项目而言非常方便。

除了使用 HapMap 镜像外，Tagger 在线服务器还能上传“Ped”格式的文件或上传从 HapMap 下载的基因分型数据。用 Nick Patterson 编写的强调程序能够自动分相这些格式的文件。该程序是基于期望值最大算法和分割捆绑方法编写的。其考虑了样本亲缘关系（如 YRI 或 CEU 的核心家系）。此外有连锁相的单体型文件也可被上传。基因型或单体型数据的文件在上传时需附加一份信息文件（包含 SNP 号和在染色体上的物理位置）。注意上传文件的大小受 Tagger 服务器限制（Tagger 网页关于文件格式）。

Tagger 默认使用的最小等位基因频率为 5%：捕获频率≥5%的所有等位基因（单个 SNP）。注意稀有等位基因在群体中的实际频率可能在 5%左右波动。不可避免的，有些变异在 HapMap 中显示频率小于 5%，而在实际群体中可能大于 5%；反之亦然。

降低入选等位基因频率阈值，将增加被捕获的等位基因数目，这可能将需要更多的 tag-SNP。在权衡分析成本和分型完成情况时需考虑等位基因的频率阈值。注意 tagger 也能够根据设定的频率域值准确地定义一组需被捕获的等位基因（SNP 和（或）单体型）。

第二步设定怎样挑选 tag-SNP。在设定的 $r^2$ 水平（限定的最小系数），Tgger 挑选 tag-SNP（tag）和定义等位基因检验用于捕获定义组中所有的等位基因。当设定 $r^2$ 值为 1.0 时，将获得一组非冗余的 tag-SNP，而所有被捕获的等位基因都伴有一个明确指示的 tag-SNP。

Tagger 软件兼有配对方法的简便性（如 LD 选择的简便性，Carlson et al. 2004）和多个遗传标记（单体型）方法的潜在效率（Stram et al. 2003；Weale et al. 2003）。多遗传标记法运行时，首先选取配对 tag，接着逐一反复剔除 tag 并且用一组明确的遗传标记来取代 tag（采用任一剩余的 tag-SNP）。在设定的 $r^2$ 值水平，如果能与排除 tag 捕获到相同的等位基因，这组被预测的遗传标记就被接受。否则，被剔除的 tag 将被视为必不可少并被保留下来。除非多遗传标记法采用有效的单体型代替 tag，否则它必须找到一组自由度为 1 的关联检验，因而需要更少的 tag 用于基因分型。这些被预测的多遗传标记被清楚地记录下来，以便在关联分析中使用（除单遗传标记检验）。

使用 Tagger 时必须选定是使用配对法还是多遗传标记法。多遗传标记法更加有效（提高挑选效率 15%～35%，这取决于基因组局部的 LD 结构和程度）。然而，由于必须完成多个标记的检验，所以多遗传标记法与配对法相比需要更准确和更完整的分型数据。虽然多遗传标记中连锁相不清楚时对于具有强 LD 的多遗传标记影响不大，但仍能干扰 Tag 的挑选。

Tagger 支持使用“包含 tag”功能指定一组 SNP。当某一区域内需要挑选更多的 tag（除了已经分型的 SNP）或某些 SNP 非常重要时，需要直接设定成 tag（通常是有功能的数据或是已经报道的关联 SNP）时，就可用到此功能。注意认为指定的 tag 只有被 HapMap 收录，Tagger 才识别（否则它不能通过 LD 来计算这些 Tag 所捕获的等位基因）。Tagger 同时也支持设定某些 SNP 不作为 tag；如排除那些可能导致错误分型的 SNPs。此外，可以给每个 SNP 赋予一个所谓的“设计分数”，Tagger 高分值的 SNP 将得到优先挑选。设定分数域值能排除分值低的 SNP 被选作 tag（“0”表示没有 SNP 被排除）。

## tag-SNP 的评价

上面部分介绍了怎样从一开始挑选 tag-SNP 或在指定的区域挑选额外的 tag。然而，Tagger 也能够使用一组 SNP 来捕获某区段内已经发现的遗传变异。要使用此功能，要求上传一个包含 SNP 参考号的“包含 tag”文件（单击“评价”框，但不挑选额外 tag）。通常，由于挑选的 tag-SNP 分型失败，这时可以用这个功能重新评价该研究所采用的工作 SNP 如何捕获那些尚未分型的 SNP。近来我们用此功能评价公司公开的全基因组分型结果在染色体上覆盖情况（Pe'er et al. 2006）。

下面的例子显示怎样使用 Affymetrix 500K 基因芯片的分型结果来评价 TCF7L2 基因内常见变异覆盖情况。TCF7L2 位于 10 号染色体，物理位置从 114 700200～114 916057bp（采用 NCBI 第 35 版物理图谱）。其与 2 型糖尿病的发病风险相关联（Grant et al. 2006），该结果在几个不同的研究中得到非常一致的重复。此外，我们想知道使用 Affymetric SNP 芯片分型的 HapMap 尼日利亚约鲁巴人群配对 LD 数据是怎样捕获 TCF7L2 内常见变异的。为了尽可能地包括启动子和其他的调控元件，我们在该基因的上、下游各延长了 30kb。然而有可能某些 Affymetrix SNP 位于该基因的外侧并与基因内的 SNP 有强 LD 结构。因为这些 SNP 有助于覆盖整个基因，所有我们需要进一步向两侧拓宽到达一个无 LD 结构的位置，如延长到上、下游 200kb 的位置。

（1）登陆 Tagger 网站（http://www. broad. mit. edu/mpg/tagger/），点击“Tagger server”链接。

（2）在“chromosomal landmarks”对话框输入染色体物理位置（就本例子输入：chr10：114470201～115146051）。

（3）从“HapMap analysis panel”下拉菜单中选择“YRI”（尼日利亚约鲁巴人群）。

（4）在“include tag SNPs”栏点击“浏览”，上传此区域内的 124 个 SNP，这些 SNP 数据来自 Affymetrix 数据库。

（5）点击“only evaluate these SNPs”对话框。

（6）在“Tagger mode”栏挑选“pairwise”框。

（7）使用默认参数点击提交（如最小等位基因频率为 5%）。

一旦 Tagger 处理完后，结果页将显示有 114 个 tag-SNP（尼日利亚群体中有 10 个 SNP 或许没有多态性，或者没能通过第 21 版 HapMap 数据的质量检测）。531 个 SNP 的等位基因频率≥5%，其中的 41%的被捕获 SNP 满足配对最大 $r^2 \geq 0.8$（平均最大 $r^2$ 为 0.57）。“检验”文件可以被下载继续分析。此文件是表格格式，里面列出了捕获每个 SNP 的检测以及两个 SNP 间的 $r^2$ 值。例如，rs12255372 被捕获时，$r^2$ 值为 0.67；捕获 rs7903146 的 $r^2$ 值较低为 0.37（在研究中这两个 SNP 与 2 型糖尿病关联性最强，提示 TCF7L2 为疾病的风险基因）。

不改变参数，使用“aggressive”（指定的多遗传标记检验）重新运行 Tagger，将有 44%的 SNP 被捕获，满足 $r^2 \geq 0.8$。例如，SNP rs7074334（MAF＝9%）被一个 Tag SNP（rs11817282）捕获，但 $r^2$ 值仅为 0.26，却有效地被 2 个 SNP 构成的单体型捕获（rs7085980 和 rs7897837），它们之间存在很好的 LD 关系，$r^2 = 0.91$。此结果提示怎样运用潜在的单体型获得检验效能和捕获遗传变异。

# 参考文献

Ahmadi K.R., Weale M.E., Xue Z.Y., Soranzo N., Yarnall D.P., Briley J.D., Maruyama Y., Kobayashi M., Wood N.W., Spurr N.K., et al. 2005. A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat. Genet.* **37:** 84–89.

Barrett J.C., Fry B., Maller J., and Daly M.J. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

Carlson C.S., Eberle M.A., Rieder M.J., Yi Q., Kruglyak L., and Nickerson D.A. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74:** 106–120.

Conrad D.F., Jakobsson M., Coop G., Wen X., Wall J.D., Rosenberg N.A., and Pritchard J.K. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* **38:** 1251–1260.

de Bakker P.I.W., Graham R.R., Altshuler D., Henderson B.E., and Haiman C.A. 2006a. Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. *Pac. Symp. Biocomput.* **2006:** 478–486.

de Bakker P.I.W., Yelensky R., Pe'er I., Gabriel S.B., Daly M.J., and Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* **37:** 1217–1223.

de Bakker P.I.W., Burtt N.P., Graham R.R., Guiducci C., Yelensky R., Drake J.A., Bersaglieri T., Penney K.L., Butler J., Young S., et al. 2006b. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38:** 1298–1303.

Gonzalez-Neira A., Ke X., Lao O., Calafell F., Navarro A., Comas D., Cann H., Bumpstead S., Ghori J., Hunt S., et al. 2006. The portability of tagSNPs across populations: A worldwide survey. *Genome Res.* **16:** 323–330.

Grant S.F., Thorleifsson G., Reynisdottir I., Benediktsson R., Manolescu A., Sainz J., Helgason A., Stefansson H., Emilsson V., Helgadottir A., et al. 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38:** 320–323.

Huang W., He Y., Wang H., Wang Y., Liu Y., Wang Y., Chu X., Wang Y., Xu L., Shen Y., et al. 2006. Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc. Natl. Acad. Sci.* **103:** 1418–1421.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Montpetit A., Nelis M., Laflamme P., Magi R., Ke X., Remm M., Cardon L., Hudson T.J., and Metspalu A. 2006. An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet.* **2:** e27.

Mueller J.C., Lohmussaar E., Magi R., Remm M., Bettecken T., Lichtner P., Biskup S., Illig T., Pfeufer A., Luedemann J., et al. 2005. Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am. J. Hum. Genet.* **76:** 387–398.

Nejentsev S., Godfrey L., Snook H., Rance H., Nutland S., Walker N.M., Lam A.C., Guja C., Ionescu-Tirgoviste C., Undlien D.E., et al. 2004. Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. *Hum. Mol. Genet.* **13:** 1633–1639.

Pe'er I., de Bakker P.I.W., Maller J., Yelensky R., Altshuler D., and Daly M.J. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.* **38:** 663–667.

Ramirez-Soriano A., Lao O., Soldevila M., Calafell F., Bertranpetit J., and Comas D. 2005. Haplotype tagging efficiency in worldwide populations in CTLA4 gene. *Genes Immun.* **6:** 646–657.

Ribas G., Gonzalez-Neira A., Salas A., Milne R.L., Vega A., Carracedo B., Gonzalez E., Barroso E., Fernandez L.P., Yankilevich P., et al. 2006. Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Hum. Genet.* **118:** 669–679.

Stankovich J., Cox C.J., Tan R.B., Montgomery D.S., Huxtable S.J., Rubio J.P., Ehm M.G., Johnson L., Butzkueven H., Kilpatrick T.J., et al. 2006. On the utility of data from the International HapMap Project for Australian association studies. *Hum. Genet.* **119:** 220–222.

Stram D.O., Haiman C.A., Hirschhorn J.N., Altshuler D., Kolonel L.N., Henderson B.E., and Pike M.C. 2003. Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55:** 27–36.

Weale M.E., Depondt C., Macdonald S.J., Smith A., Lai P.S., Shorvon S.D., Wood N.W., and Goldstein D.B. 2003. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73:** 551–565.

# 互联网资源

**http://www.broad.mit.edu/mpg/haploview/** Haploview is a useful tool for extracting genotype data and running tag-SNP selection.

**http://www.broad.mit.edu/mpg/tagger/** Tagger can be used for tag-SNP selection and evaluation using HapMap data.

# 21 Haploview：可视化和分析 SNP 基因型数据

Jeffrey C. Barrett

*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom*



## 简介

关联研究包括评价、剖析、产生和分析大量数据，完成整个过程经常需要几个月的时间。遗传变异的大规模测量项目，如 HapMap 计划（Altshuler et al. 2005）和迅猛增加的 SNP 分型数据已经为关联研究提供了许多振奋人心的机会。然而数据的混杂和分析数据产生的困难越来越令人困扰。Haploview 由 Broad 研究所的 Mark Daly 实验室开发，归纳每日常规的分析活动以形成一个易于使用的软件包。此软是用 Java 语言编写，原代码免费开放，其能在 Windows、MacOS 和 UNIX 等操作系统上使用。该软件，文件和示例（包括在本章中运用的例子）都能从 http://www.broad.mit.edu/mpg/haploview/网页获得。Haploview 的许多特征都适合关联分析。其中这些特点通过下面这个关联分析例子进行演示。使用 Haploview 可以完成下面分析：①分析 HapMap 数据和选择 tag-SNP，②评价疾病基因分型数据的质量，③关联检验，④为深入研究疾病关联基因评价染色体区域。

## 分析 HAPMAP 数据

国际人类基因组单体型图计划（Altshuler et al. 2005）致力于在全基因组内大规

模的发现和分型 SNP（在 270 个样本中已经分型 400 多万个 SNP），以便阐明整个基因组的变异。Haploview 与 HapMap 网站完全整合，能上传从 HapMap 下载的基因分型文件(http://www.hapmap.org/)或直接在线进入 HapMap 数据库。HapMap 支持输入染色体物理位置，基因名或某一具体的 SNP 进行搜索。

## 可视化连锁不平衡

对连锁不平衡传递模式（LD）的逐渐认识（Daly et al. 2001；Phillips et al. 2003；Hinds et al. 2005）使得研究人员能够在设计和分析关联研究方面提出许多重要的改进。基因组的大部分显示出强 LD 的“块状”结构，它们之间被明显的断裂点隔开，而这些断裂点来源于进化上的重组热点（Myers et al. 2005）。图 21-1 显示 HapMap 里 TCF7L2 内去掉等位基因频率小于 5% 的 SNP 后的 LD 结构（本章以 TCF7L2 为例）。使用者可以定义 LD 的颜色，空间以及其他参数细节。基因和 SNP 的位置信息能够直接在线下载，显示在图表的顶部。LD 模块被计算后用黑线标示。虽然模块的边界不固定，但仍很接近。它们能通过定义模块或延 LD 图顶部通过拖拽的方法手动完成模块的定义。

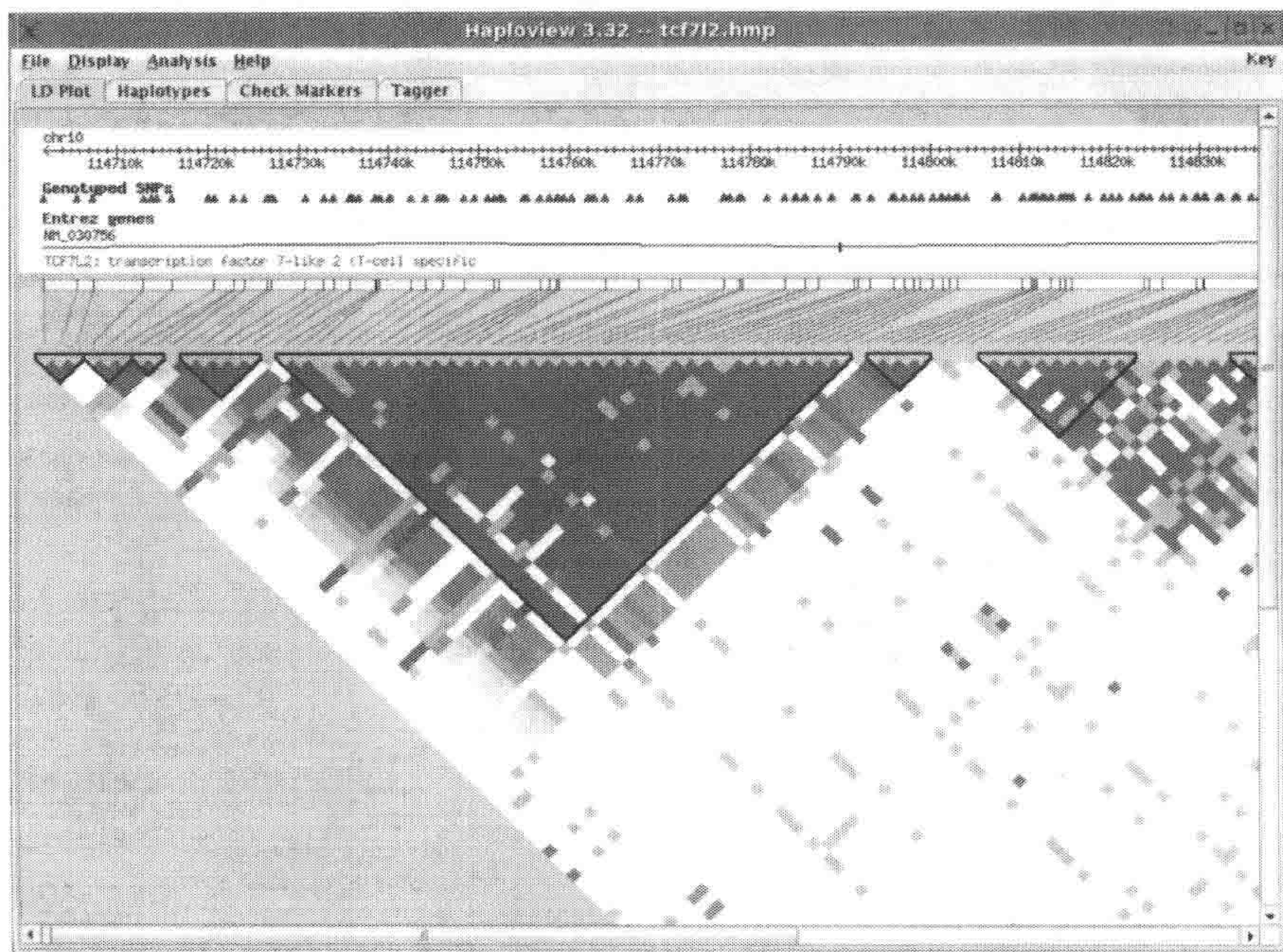



图 21-1　Haploview LD 图显示来源于 HapMap 的犹他州的 90 个欧洲后裔 TCF7L2 分型数据的 LD 图。在图顶部有一条标记有染色体物理位置，SNP 以及基因位置的参考线。红色表示 SNP 间存在强 LD，白色为没有 LD 结构，缺乏统计学意义则用浅蓝色表示。基因组中经常可以看到存在 LD 的“块状”三角形区域，大三角形表示此区域高度 LD。三角形之间被一些狭窄区域分开，而在这些区域内即使是相邻 SNP 的都是完全独立的（见图版）

## 选择 tag-SNP

HapMap 提供的挑选 Tag-SNP 的功能可能是使用 LD 知识最多的情况。因为强 LD 的特点是相邻 SNP 是高度冗余的，所以使用少数的 SNP 就能覆盖该区域内的所有变异信息。Haploview 支持使用 Paul de Bakker's Tagger 算法挑选 SNP（de Bakker et al. 2005）。在多数情况下，Haploview 在一组数据中评价所有 SNP 的 LD 结构，并选择能正确捕获所有变异的一组最小 SNP。tag 挑选的过程可以通过几种方法来完成，如考虑 SNP 的“设定值”，准确设定纳入 tag 的 LD 域值，或指定多个 SNP 作为 tag 的方法。更详细的信息参见第 20 章 Tagger 运算法则的描述。

使用默认的次要等位基因频率参数筛选 TCF7L2 数据，然后运行 Tagger，结果显示大约 50 个 SNP 可以捕获 TCF7L2 内的所有常见变异。图 21-2 显示每个 tag 捕获的 SNP。表中的数据结构是一种常规模式：几个“最佳”（如最大信息量）的 tag 捕获成簇的常见变异，而剩下的 tag 则只能捕捉自己（因此它们不是高效 tag）（Barrett and Cardon 2006）。因为用于关联研究的基因分型数据经常受到样本资源的限制，所以重新运

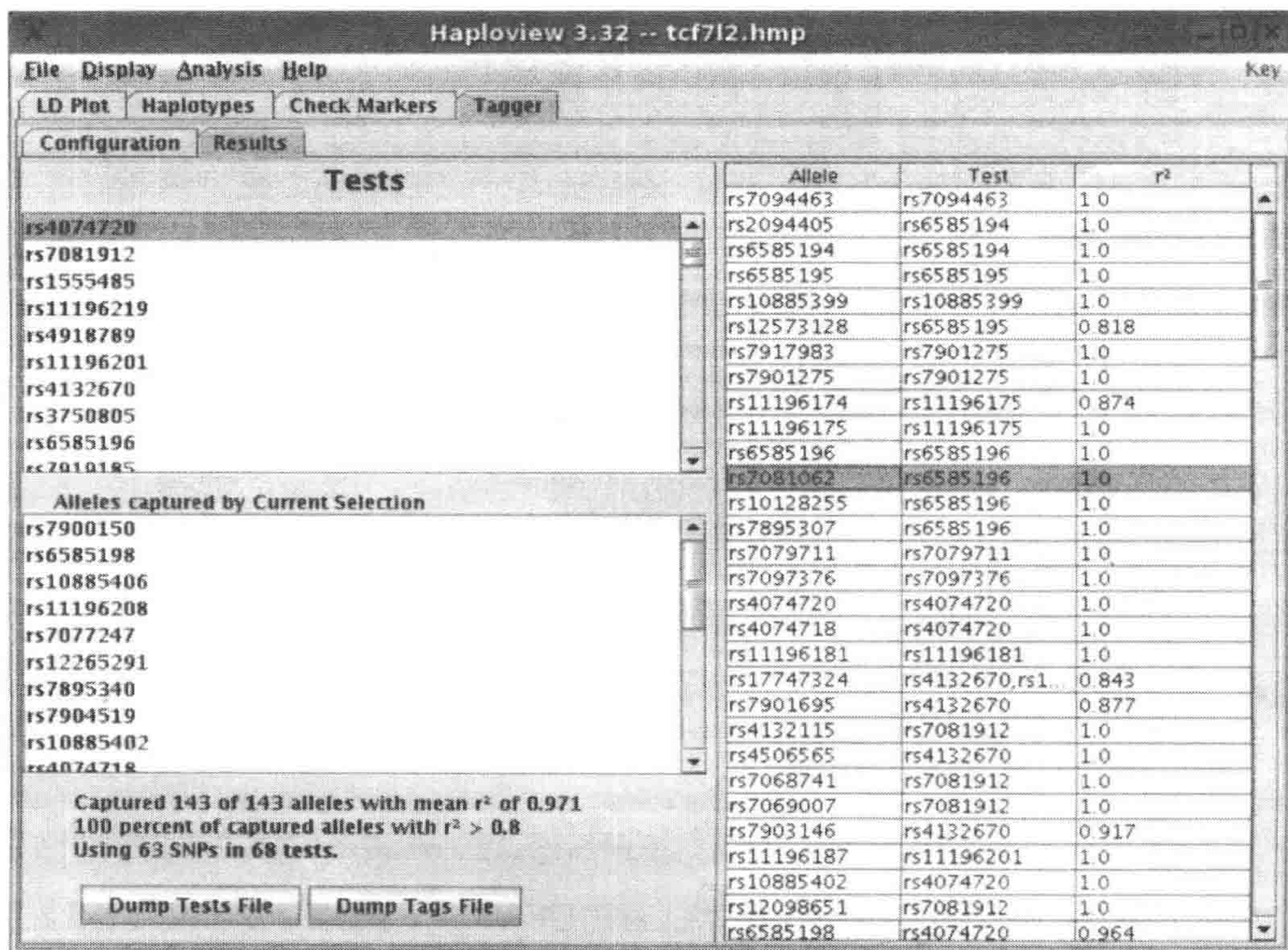



图 21-2　Haploview 中 Tagger 功能选择 HapMap TCF7L2 里 tag 的界面。tag-SNP 被列在图表的左上框，当前光标处的 SNP（rs4074720）所捕获的 SNP 被列在左下方的对话框里。右侧表中列出了所有被检测的 SNP 及其他最佳的 tag。在图表的左边底部的两个按钮是用来保存和输出检测结果

行 Tagger 能够获得最佳的一组 tag。例如，仅需要 10 个 tag 就能捕获一半的 TCF7L2 变异。Haploview 能输出一个“tag”文件，里面列有需要分型的 tag 和一个“test”文件，其包括了在关联研究中需使用到的 tag 及其构成的单体型检验结果（第 20 章关于单体型检验部分）。

## 关联分析研究

Haploview 的第二个重要部分处理用于关联研究的基因分型数据。Haploview 提供了一套完整的分析工具来控制分型数据的质量，检测 LD 结构和单体型模式，以便进行关联分析。在 TCF7L2 的研究中，首先从 HapMap 的数据库中挑选 10 个 tag-SNP，然后模拟结果用于指导 500 名病例对照的关联研究。

## 数据质控

虽然，随着 SNP 基因分型更加自动化和可靠，在技术上取得了进步，但是进行分析时，数据的质量控制仍然是关键的一步。Haploview 计算了许多测量质量控制的指标，可以用来检测 SNP 基因分型和 DNA 样本的质量。

等位基因分型率：大量的缺失数据（来源于 SNP 分型或样本信息）提示分型结果质量不好或样本信息不全。这些 SNP 和个体将从分析中被去掉，以免人为引入误差。

哈迪-温伯格平衡：偏离哈迪-温伯格平衡提示该 SNP 被定位到基因组的多个区域，或基因分型结果不好（如不能正确反映杂合与情况）。

孟德尔遗传错误：对于家系样本，Haploview 计算从亲代到子代的传递错误，同时也提示有分型错误。

次要等位基因频率：虽然就次要等位基因频率而言并不属于质量检测，但筛选 SNP 的频率有助于后续分析。

可以用 Haploview 默认的参数和手动调节参数来筛选 SNP。同时也可以手动直接纳入或排除 SNP 标志，替代自动筛选。

## 检验关联

Haploview 的特点是能够做家系和病例对照的单个 SNP 和单体型的关联分析。结果给出疾病关联的等位基因、病例对照的等位基因频率、$X^2$ 检验的统计值以及评价显著性的 $P$ 值。虽然其他的方法也能够完成关联分析（如线性回归模型），但用 Haploview 做基本的关联分析还是不错的选择。

上传从 HapMap 下载包含分型数据的“检验”文件，Haploview 被设置自动进行 SNP 和单体型的检验。Haploview 利用 EM 算法（Qin et al. 2002）计算出单体型连锁相用于人群频率（“单体型”表）和关联检验。TCF7L2 的数据分析结果显示 SNP rs10509969 与糖尿病强关联，这结果值得深入研究。

由于关联研究涉及多个变异和单体型的检测，所以在评价显著性时应考虑这些检测。通过排列处理 Haploview 产生经验的显著性结果（如不管总的检测次数）。先在个体中排列病例对照，然后运行常规的关联检验。通过多次重复此过程，Haploview 建立一个零分布的检验统计，在此分布中的结果属于偶然事件。实际关联检验的显著性比随机结果大许多倍，其比预期观察到的关联结果更显著。用排列模拟关联研究显示排列后检验出的关联性仍然显著（图 21-3）。

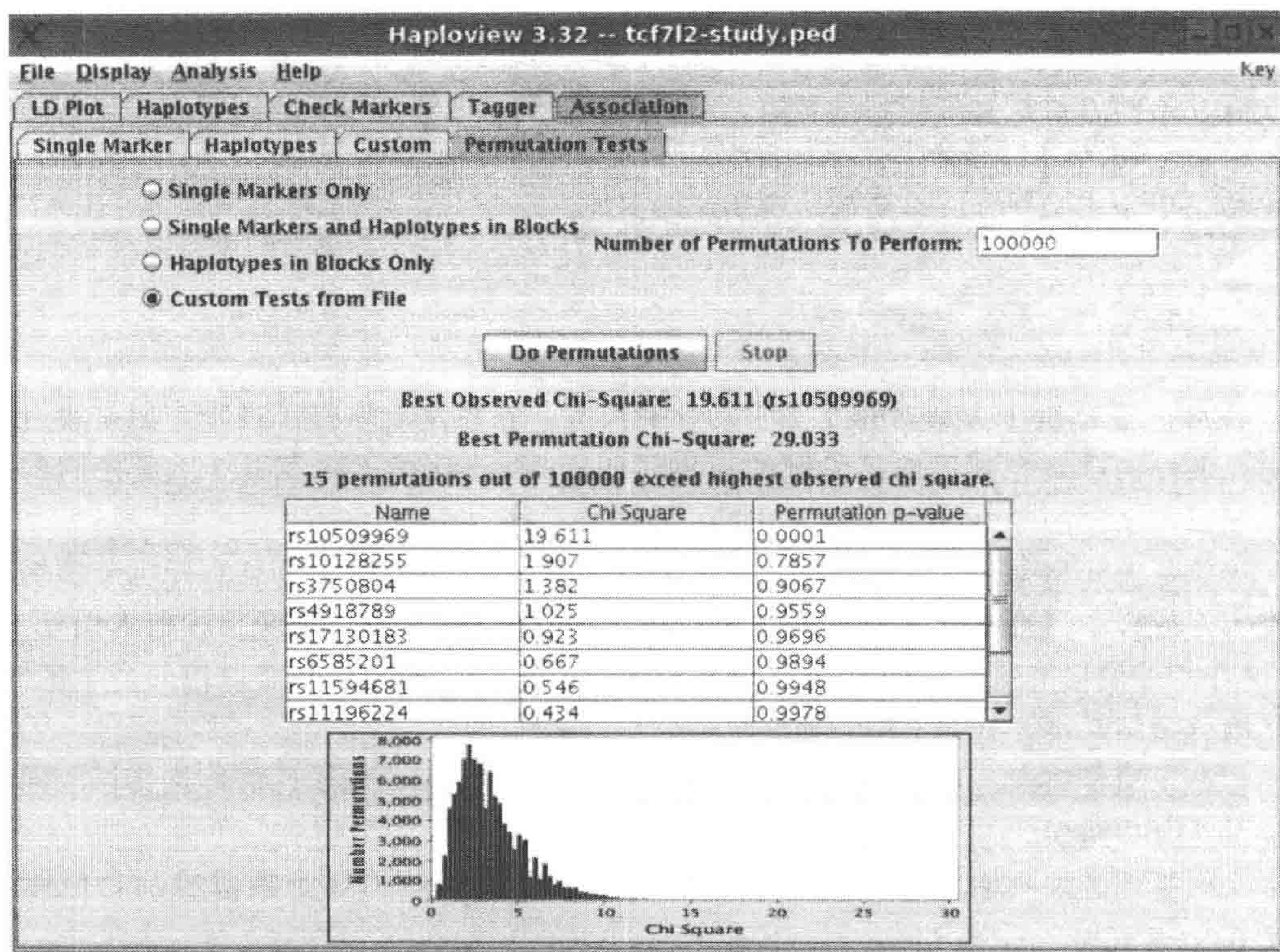



图 21-3 Haploview 模拟 500 名病例对照关联研究的排列检验界面。经过 100 000 次排列检验后，显示 SNP rs10509969 与糖尿病强关联，*P* 值 0.001 设定为经验的显著性水平，用来评价实验结果的显著性

## 接下来的工作

一旦 Haploview 检验到关联，接下来的工作就是深入剖析此结果，查找致病基因。返回重新分析 HapMap 的原始数据显示此关联性的 SNP 存在于一个 12kb 的强 LD 模块内。接下来研究的方法包括分型此 LD 模块内的其他 SNP，在数据库中搜索此区域内已知的无义突变，以及直接对此区域进行测序。

## 总结

将低成本、高通量的基因分型与常见变异的数据库，如 HapMap 结合起来不仅可以生成大规模的数据，而且为发现复杂疾病的致病基因提供了研究前景。将原始数据转换成生物学的信息要求有一个简明界面的分析工具来处理重要的分析工作。Haploview 是一个有用的生物学软件，可以用于处理多种问题，如挑选 Tag-SNP、评价 LD 情况，以及关联检验。一个综合性的研究需要采用多个软件尽可能多地从原始数据里收集信息，而 Haploview 为研究大多数常见遗传问题提供了一个好的切入点。

## 致谢

Haploview 由 Jeffrey Barrett、Julian Maller、David Bender 以及 Mark Daly 开发和维护。感谢 B. Herrera 和 M. Lincoln 对本文的评阅和指正。

## 参考文献

Altshuler D., Brooks L.D., Chakravarti A., Collins F.S., Daly M.J., and Donnelly P. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Barrett J.C. and Cardon L.R. 2006. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38:** 659–662.

Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J., and Lander E.S. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29:** 229–232.

de Bakker P.I.W., Yelensky R., Pe'er I., Gabriel S.B., Daly M.J., and Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* **37:** 1217–1223.

Hinds D.A., Stuve L.L., Nilsen G.B., Halperin E., Eskin E., Ballinger D.G., Frazer K.A., and Cox D.R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.

Myers S., Bottolo L., Freeman C., McVean G., and Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310:** 321–324.

Phillips M.S., Lawrence R., Sachidanandam R., Morris A.P., Balding D.J., Donaldson M.A., Studebaker J.F., Ankener W.M., Alfisi S.V., Kuo F.S., et al. 2003. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* **33:** 382–387.

Qin Z.S., Niu T., and Liu J.S. 2002. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71:** 1242–1247.

## 互联网资源

**http://www.broad.mit.edu/mpg/haploview/** Haploview program home page, Broad Institute of MIT and Harvard. Haploview bundles many everyday analysis tasks into one easy-to-use package. It is an open-source program written in Java and capable of running on Windows, MacOS, and UNIX platforms.

**http://www.hapmap.org/** International HapMap Project home page. The HapMap project is a major repository for genotype data providing remarkably detailed data sets in three geographically discrete human subpopulations: Europeans, Africans, and Asians. Tools built upon this resource are currently the state of the art for those regions of the genome that have not been targeted for resequencing.

# 22 FFPE 样本进行拷贝数分析时需要考虑的问题

Sharoni Jacobs

*Affymetrix, Santa Clara, California 95051*



## 简介

### WGSA、Mapping 微阵列和拷贝数检测

全基因组样品分析（whole genome sample analysis，WGSA）能同时对上千个 SNP 标记，通过 Affymetrix GeneChip Mapping 微阵列进行基因分型（Kennedy et al. 2003）。使用相同的 Mapping 微阵列芯片数据，可以对每个 SNP 进行拷贝数（CN）变化的定量分析（Bignell et al. 2004；Huang 2004；Lin et al. 2004；Nannya et al. 2005）。这样，一次实验中，Mapping 500 K 微阵列芯片可以提供超过 500 000 个 SNP 的基因型和拷贝数信息，其中包含 Mapping 250K Nsp 微阵列以及 Mapping 250K Sty 微阵列。

Mapping 微阵列由短的寡核苷酸（25-mer）探针组成，这些探针可以特异性的与某一

SNP 的不同等位基因发生退火。在 WGSA 过程中，DNA 在加入目标微阵列前，先使用限制性酶进行消化（对于 Mapping 500K 微阵列使用 NspI 或 StyI)。消化末端连接为短的寡核苷酸作为其后续 PCR 步骤时引物的退火模板。在 PCR 期间，根据消化后的 DNA 片段大小情况，一部分基因组被选择性的进行扩增为特定的片段——仅有长度范围为 100～1100bp 的片段被扩增。这之后，PCR 的扩增子经 DNase I 消化而片段化，生物素标记后，与微阵列进行杂交。芯片用荧光标记的 streptavidin 进行染色，对荧光探针的强度进行测定。对测定的强度值进行比较，以确定芯片上的每个 SNP 的基因型。

某个 SNP 位置的拷贝数也可以根据 SNP 探针的强度值进行确定（Huang　2004)。当某 SNP 为杂合子基因型时，标记后的 DNA 杂交至不同等位基因特异性的探针后，通过单独的强度测定，能够识别等位基因特异的拷贝数（LaFramboise et al.　2005；Huang et al. 2006)。等位基因特异性的拷贝数检测的优点在于能够对染色体发生加倍的区域进行识别，但同源区域如果被删除则不可进行识别（这种情况下，不会发生整体的拷贝数的改变。如图 22-1 中 p-末端发生的情况)。当 DNA 确实发生整体上拷贝数的改变时，等位基因特异性的拷贝数可以为区别不同类型的染色体改变提供有价值的信息提示，比如两种等位基因都增加，一个等位基因增加（如图 22-1 中 q-末端发生的情况)，以及获得一个等位基因同时伴随丢失另一个等位基因（如图 22-1 中“3＋0”区域中发生的情况)。

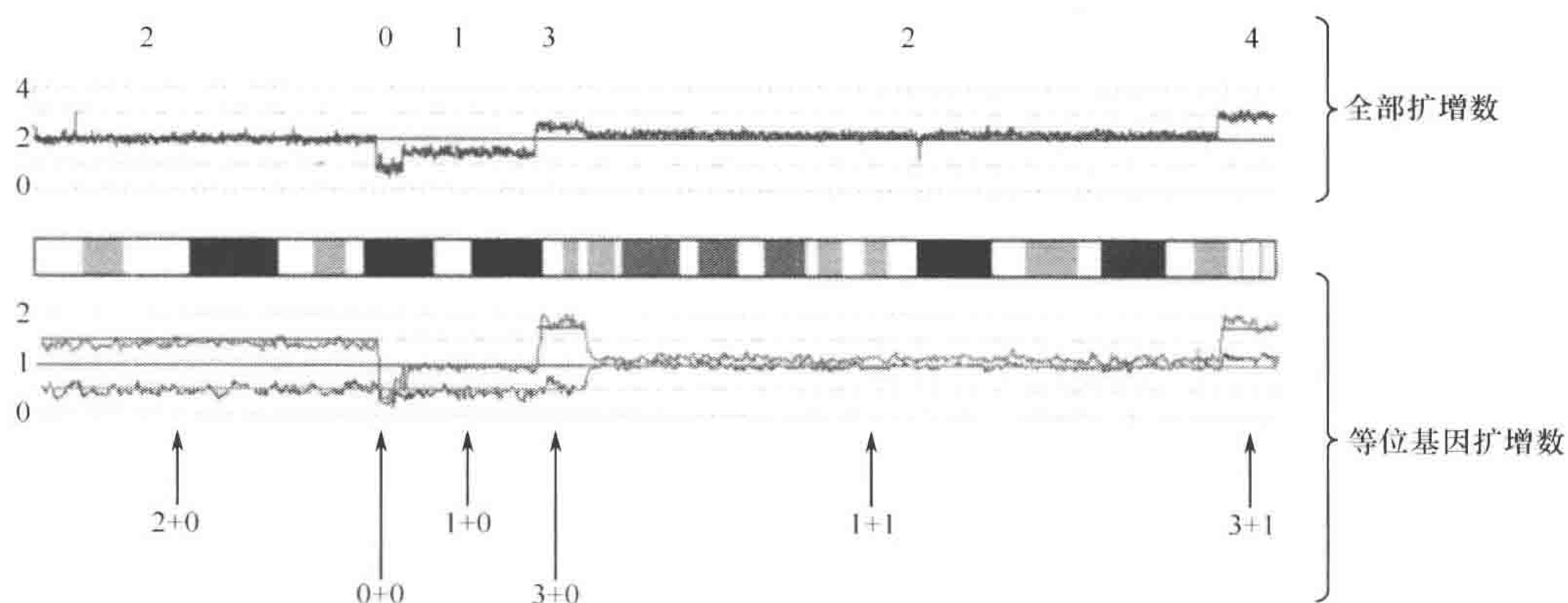



图 22-1　CNAG 中单条染色体中拷贝数的图像。整体的拷贝数预测在上面部分用蓝色标出，平均有 10 个 SNP。下面部分为等位基因特异性的拷贝数，用红色和绿色线标出，分别代表高丰度和低丰度的等位基因。黑色的数字代表每个区块中整体的拷贝数，红色和绿色的数字代表这些区域中等位基因特异性的拷贝数。(见图版)

## 使用 WGSA 进行 DNA 含量的定量分析

拷贝数检测可以对 DNA 的拷贝数提供定量的评价。因此，WGSA 不会歪曲 DNA 片段的相对代表性，这一点是很重要的。消化、连接以及片段化等步骤都不能偏向于某一特定的 DNA 区域。另外，生物素标记使用末端脱氧转移酶（TDT）作为末端标记反应，这样保存了成比例的 DNA 片段的代表性。

另外，PCR 作为最高效的改变 DNA 片段相对含量的方法，其终止于线性扩增范围

内。观察到的 Mapping 500K 微阵列的结果表明，当将 WGSA 用于高质量的 DNA 片段时，对于 DNA 片段的相对含量只有很小的影响，并且大多数拷贝数分析工具甚至能够忽略这种最小影响的存在（Huang　2004；Lin et al.　2004）。

### FFPE 和拷贝数检测

使用福尔马林固定后石蜡包埋（FFPE）的样本作为 DNA 的来源时，需要对样本的处理和数据分析进行特殊的考虑。这些样本常常发生降解，不同样本之间降解的程度区别很大；因此，一些 FFPE 资源的 DNA 降解的特别严重（能提供的链长不超过 100bp），但是另一些 FFPE 样本可以提供相对完整的 DNA。除了降解的程度不同外，DNA 也可能发生化学修饰。最终，FFPE 的 DNA 样本常常包含残存物的污染，其能够抑制下游的分子反应。使用 FFPE 样本用于 Mapping 芯片的操作指南在第 16 章中进行了描述。

WGSA 的步骤中，最容易受 FFPE DNA 质量影响的是 PCR 的环节。正如在简介中提到的那样，在 PCR 环节中，高质量的 DNA 样本可被扩增至最大为 1100bp 的片段。但是，发生降解的 FFPE 样本常常不能提供这样长度的初始模板。因此，PCR 之后，较大的扩增子以及相关联的 SNP 可能发生缺失。另外，污染物的存在和 DNA 的修饰可能影响反应，降低扩增的效率并且导致片段大小发生偏差，小片段扩增的情况优于更大的片段。

片段大小的偏差可以造成一种局面，除真实存在染色体异常的情况外，小片段将可能提示拷贝数目的增加，而大片段可能被提示存在缺失（图 22-2 第一行）。借助计算机可以对此片段的偏差进行校正，对于大片段扩增子的扩增失败进行接纳，排除其不进行分析。本章中，我们对于从降解和污染的 FFPE 样本中获得可靠的拷贝数预测所需要的步骤进行探讨。

## 需要的软件

对于 Affymetrix Mapping Array 数据的拷贝数分析有许多的软件包可以使用。它们包括 Affymetrix Copy Number Tool（CNAT；Huang　2004），在 Affymetrix 的网站上可以免费获得（www. affYmetrix. com）；学术免费软件，如 dChipSNP（Lin et al.　2004），CNAG（Nannya et al.　2005）、GEMCA（Ishikawa et al.　2005；Komura et al.　2006）；商业化的软件工具，如 Partek Genomics Suite、Sapio Science's Exemplar for Copy Number，以及 Stratagene's ArrayAssist Copy Number。当使用高质量的 DNA 时，这些工具中都可以用于对拷贝数的检测。撰写本文时，只有 CNAT 4.0 和 CNAG2.0 适用于分析 FFPE 样本的分析；因此，本章节终点关注的是这方面的应用（更早版本的 CNAT 应该不能使用于这个领域）。将来，其他工具将不断地被发展和升级，最后也将能用于 FFPE 来源样本的分析。

为了对 FFPE 样本进行正确的分析，所使用的软件工具需要具有两个特征（Jacobs et al.　2007）。首先是要能够对片段长度的偏差进行补偿校正。这可以由二次回归完成，这在 CNAT 和 CNAG 中都是可以自动完成的。其次就是要能够根据片段的大小过滤出 SNPs。过滤可以手动输入片段大小的阈值而完成。目前，大多数的应用程序都是允许使用者设定片段大小的阈值，但是仅仅 CNAT 和 CNAG 能对片段大小的偏差进行校正。

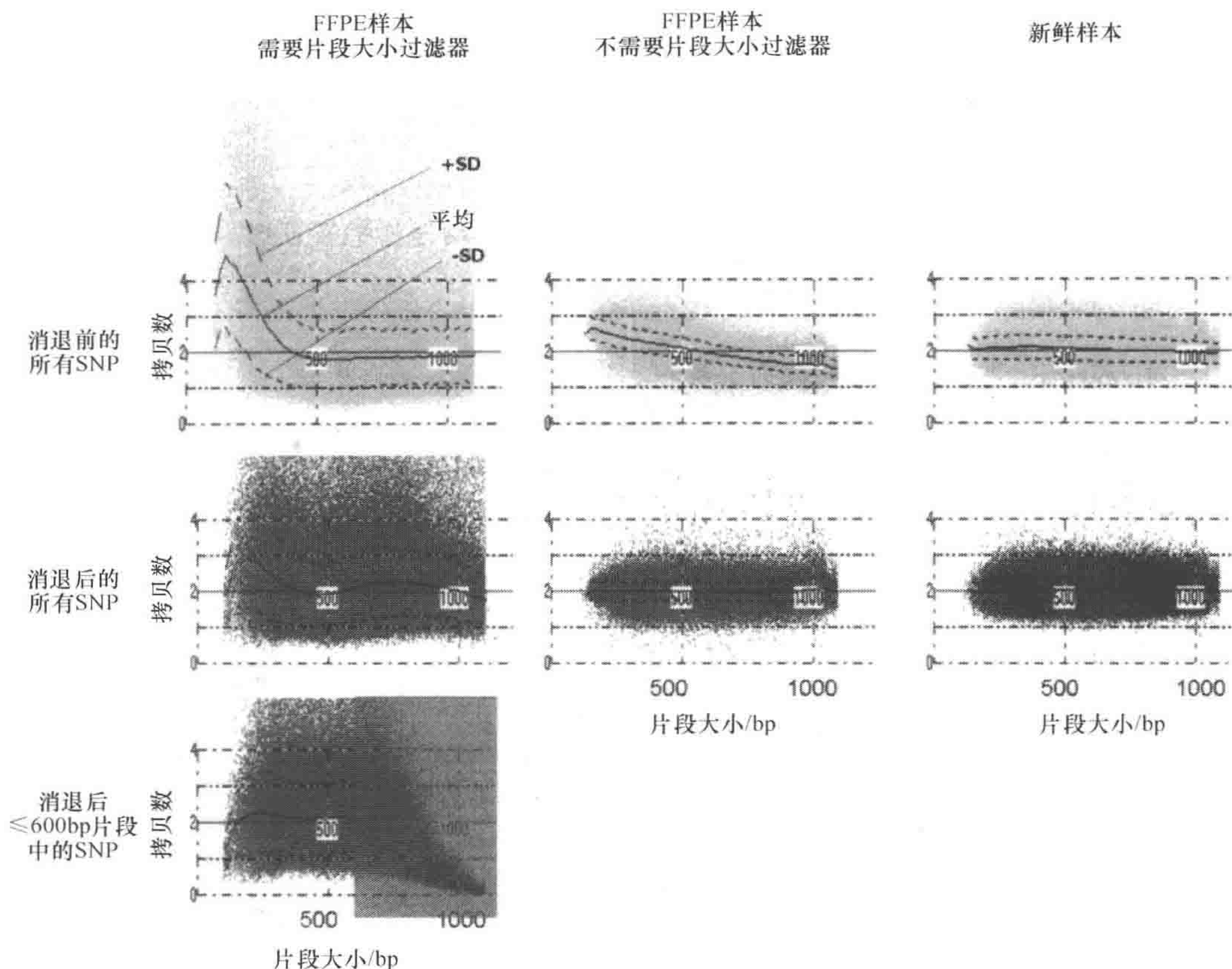



图 22-2 CN对片段大小作图。按照芯片上每个SNP的拷贝数和片段大小绘出此图。第一行中，显示的是原始的拷贝数的情况：实线代表平均拷贝数；两根虚线表示标准差。在FFPE样本中，平均拷贝数有一定的波动，但是在新鲜样本中，波动的程度就很小。中间一行通过二次回归对片段产生的偏差进行了校正。FFPE的样本中，有一组（中间一列）无论片段大小如何平均拷贝数都很一致，因而提示其不需要进行片段大小的过滤。回归校正不足去除另一组FFPE样本的偏差（左边一列）。对于这个例子中，排除位于＞600bp片段上的SNP，并且结合回归校正，可以去除大部分由于片段大小对分析中纳入的SNP的拷贝数预测带来的偏差（最后一行）

# 关于参照所考虑的问题

## 参照类型、数量和性别的选择

为了定量地确定一个样本中存在DNA的拷贝数的情况，需要有一个已知相对倍数的参照。这些样本的来源应该为“正常的”，无疾病的个体（如HapMap中的样本）；癌症患者的非肿瘤组织；或者先天性疾病的子女的未发病的父母。参照可以是单独的一个DNA样本（如在肿瘤-正常配对分析中），或者多个DNA样本的组合。当使用CNAT用于非配对的分析时，推荐最少需要25个参照以获得最高的信噪比。对于

CNAG，当手动选择参照时，建议最少需要有 5 个参照被选择。

软件程序假设参照样本是二倍体。因此，当 X 染色体要进行分析前，参照的性别应该只有一种，最好是女性。如果男性样本作为参照，那么应该注意到两个拷贝的假设对于 X 染色体不能成立，因而对于这条染色体拷贝数的解释必须进行调整。

当没有参照时（或者由于参照库的增补时），HapMap 的 48 个样本经 Mapping 500K Set 的数据结果可以使用，可以从 Affymetrix 的网站上进行下载。

### 批次效应

对于最高信噪比的拷贝数数据，参照必须与病例样本来自相同一个实验室，有相同的使用者，并且在同一批次里。由于杂交时间、洗涤次数以及其他的因素导致的批次效应可能导致对于拷贝数的预测增加少量的噪音，但是其程度相对是很小的（与表达的数据相比，其容易显著性的受不同批次的影响）。因此，当没有相同批次的参照时，其他的参照也可以进行采用。

### 配对和非配对分析

对待查 DNA 与匹配参照样本比较进行配对分析是可行的。体细胞改变的识别可以根据对同一患者的肿瘤组织和非肿瘤组织进行比较而实现。对于生殖细胞的突变，孩子的 DNA 可以和父母的 DNA 进行比较，以探测新发的染色体异常。当不能获得来自患者的非肿瘤 DNA 或其父母的 DNA 时，可以使用非匹配的参照样本进行非配对分析。一般而言，这种情况下，将使用多个参照组成的库进行分析。CNAT 和 CNAG 都可以进行配对和非配对的分析。

### 在 CNAG 中自动选择参照

当进行非匹配的分析时，CNAG 可以自动选择最适匹配试验 DNA 样本的参照。当使用这个选项时，CNAG 将从 CNAG 文件夹中提取的每个参照样本和试验 DNA 样本相比计算出，与拷贝数预测相关联的标准差（S. D.）。另外，CNAG 将对试验 DNA 样本和多个参照进行比较，直至这个参照或这些参照进行拷贝数预测时的标准差最小。如果想确定在自动选择时哪个参照样本被选择，那么当拷贝数信息出现的时候，点击 CNAG 窗口顶部的“Info”图标。

### FFPE 和非 FFPE 参照

当分析 FFPE 样本的拷贝数时，所用的参照可以是非 FFPE 或 FFPE 的样本，但是由于下一章节中描述的原因，使用非 FFPE 的参照更为优先。

## FFPE 样本的拷贝数分析

### 对于片段大小的偏差进行补偿校正

在对 FFPE 样本进行拷贝数的预测时，常常会出现片段大小存在偏差；小片段将被

预测具有更高的拷贝数，而相对较大的片段将被预测更低的拷贝数。这种类型的偏差在拷贝数与片段大小比较图图中看到（图 22-2 第一行）。CNAT 和 CNAG 可自动进行二次回归，这样可以减少片段大小对平均拷贝数的影响（图 22-2 第二行）。由于降解的样本不能提供大片段的模板，这种补偿是不充分的，不能彻底移除片段大小对拷贝数预测的影响。这种情况下，也需要另外进行一步，排除大片段不进行分析（图 22-2 第三行）。

## 片段大小过滤筛选的应用

CNAT 和 CNAG 中都可以进行较大片段的排除。当使用 CNAG 时，在参数菜单项目中选择“SetFragmentLengthRange”，然后输入在分析中需要纳入的片段大小范围。在 CNAT 中，点击 CBAT4 批量分析窗口中的“Advanced Analysis Options”图标；点击声明之后的“Restrict Analysis to SNPs on Fragment Sizes Ranging:”检验栏；然后填写希望分析长度的最小值和最大值。

在 CNAG 中，样本进行分析后片段大小的选择还可以进行调整。在 CNAT 中，在对样本或一系列样本进行分析前需要选定样本片段大小的过滤范围。改变片段大小过滤的情况后，需要对样本重新进行分析。在 CNAT 中，单批次中多样本分析时将使用相同的片段大小过滤设定。

## 片段大小过滤筛选的选择

目前，优化片段大小过滤选择的指南只能在 CNAG 中进行。对于给定的 FFPE 样本，为了选择片段大小的过滤设定，可以按照图 22-3 标明的流程图，使用拷贝数与片段大小比较的图进行。当给定的样本的拷贝数数据显示时，可以在参数菜单中选择“Fragment Length Plot”，从而可以看到这些图。由于在标准化的步骤中，SNP 都成组聚集，因而在 PCR 过程中没有扩增的那些 SNP 可能会影响发生扩增的较小片段的拷贝数的预测。因此，这些大片段的 SNP 必须从分析中排除。为了确定纳入 SNP 的上限，需逐步地排除更多的 SNP（如首先纳入的 SNP≤1000 bp，然后纳入的 SNP≤900 bp、800 bp 等）。每次的排除将会改变拷贝数与片段大小比较图的结果，因此，一旦进行了正确的排除之后，分析中保留的片段大小中的 SNP 应提供一致的平均拷贝数穿过 $x$ 轴（图 22-3B）。这表明片段长度将不再对拷贝数的分析造成影响，对于剩下的 SNP 可以进行可靠的拷贝数的预测。一旦选择了正确的过滤设定，再次选择参数菜单中的“Fragment Length Plot”以退出拷贝数与片段大小比较图。

有的样本中，有很大比例的 SNP 需要排除，它们所提供的信息中有更高的背景噪音，这增加了对一些数据进行滤除的必要性。需要对 500bp 以下的 SNP 进行排除的样本可能不能提供任何满意的拷贝数的信息。请注意，片段大小的阈值并非直接由第 16 章中的 QC 检验进行预测。

由于 FFPE 样本 DNA 的质量变化很大，不同的样本将可能需要不同的样本大小过滤设置；因此，较差质量的样本将需要排除更多的 SNP。当在一个批次中，大量的 FF-PE 样本要进行分析时，可以使用单个片段大小的过滤设置。注意当我们在不同样本间

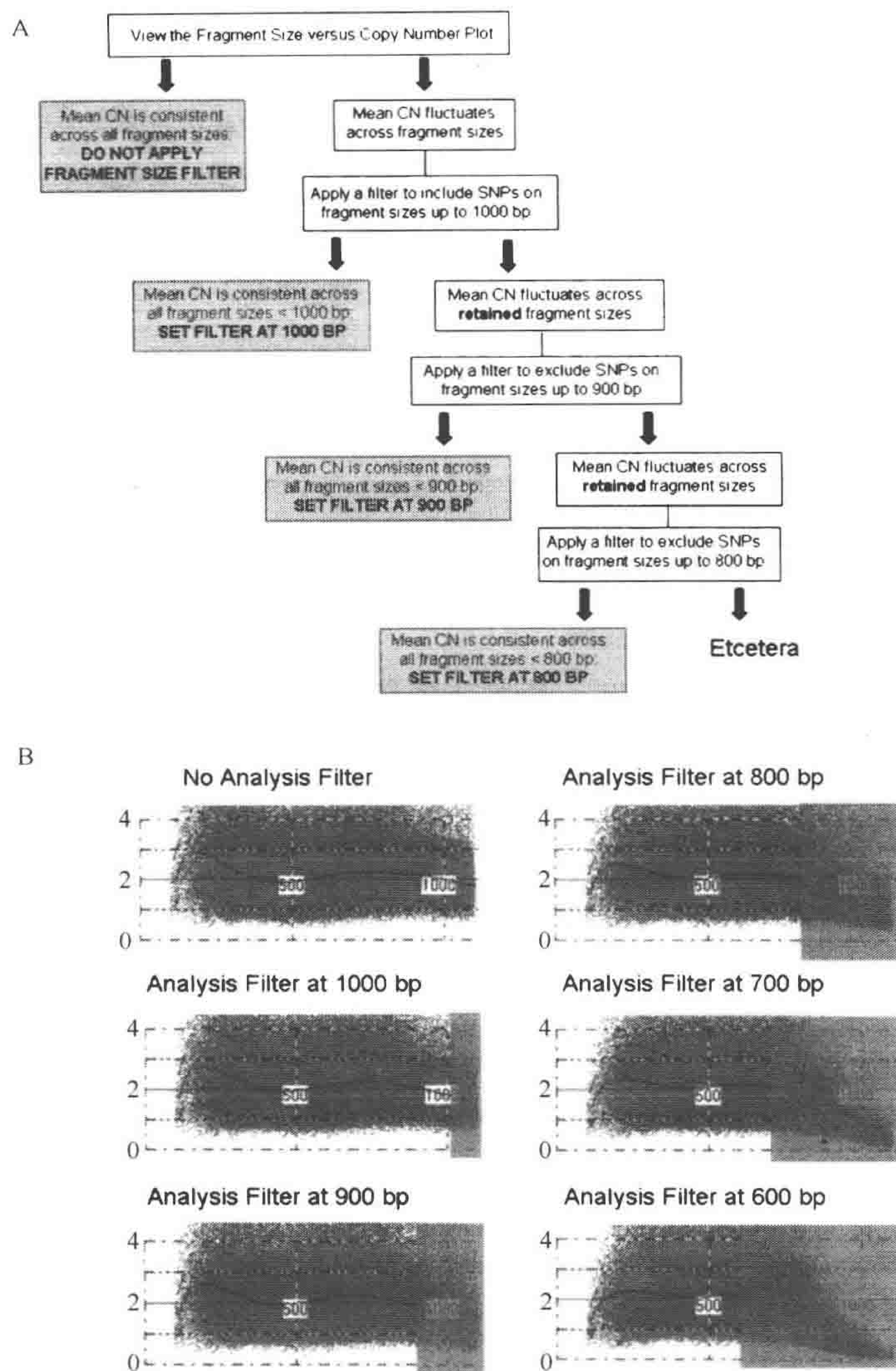



图 22-3　对于一个 FFPE 样本最佳片段大小筛选的确认。A. 这里描述的是使用 CNAG 确认对于一个给定的 FFPE 样本进行最佳片段大小筛选的步骤。B. 观察拷贝数与片段大小比较图。继续进行更多精确的片段大小筛选被用于 FFPE 样本，虽然平均拷贝数有所变动，这是由于进行筛选前以及中等长度片段过滤筛选后，仍存在片段大小的因素引起的，但是一旦按照 600～700bp 进行筛选后，平均拷贝数则稳定在接近 2 的水平。在这个例子中，600bp 将被选择作为片段大小的切点值

使用通用的过滤设置时，可能会增加较差质量样本的噪音并且降低对于较大片段 SNPs 可以提供较好信息的样本的分辨率。

## FFPE 样本作为参照

当 FFPE 样本用作拷贝数预测的参照时，片段大小的切点值的确定将可能出现问

题。此问题的发生是因为在试验的 FFPE 样本和参照 FFPE 样本进行 PCR 后，大片段中的 SNP 都可能出现同样的缺失。注意拷贝数的预测使用的是这两个样本强度的比值，前提假设是参照样本代表的是二倍体的状态，而不考虑真实的拷贝数的情况。在试验 DNA 和参照 DNA 中，大片段里的 SNP 都出现相同的减少，软件也将能预测 FFPE 样本是二倍体。由于大片段将得到一致的拷贝数为 2 的预测，且标准差很小，因而，拷贝数片段大小比较图图对于确定非信息的片段大小将不再有用。因此，虽然 FFPE 参照在分析 FFPE 样本是可以被采用，但是其确定所需的片段大小切点值的能力将大打折扣，对于来自于较大片段 SNP 的信息，其可靠性假设可能出现错误（图 22-4）。

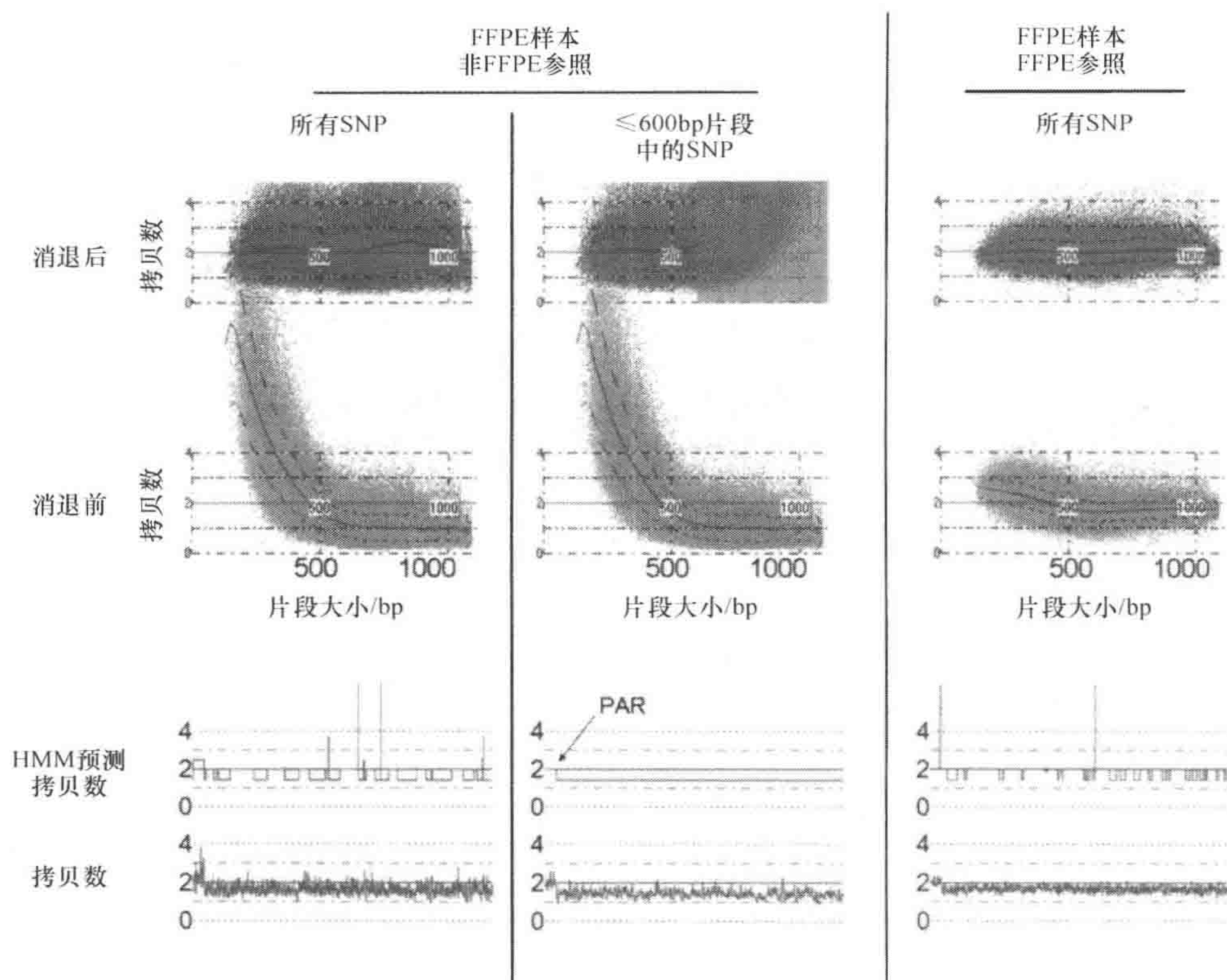



图 22-4 在检测 FFPE 样本的拷贝数时，对 FFPE 和非 FFPE 参照的比较。一个 FFPE 的样本与多个非 FFPE 的参照进行比较（左列和中列）。拷贝数片段大小比较图提示，当所有的 SNP 都纳入分析时（左列），即便进行了回归调整，片段的偏差也存在，但是当按照 600bp 中列为过滤标准进行筛选后，剩余的 SNP 进行分析时，此片段的偏差就消失了。Hidden Markov Method（HMM）预测和拷贝数预测结果都表明，此男性样本当所有 SNP 都纳入分析时，对于 X 染色体拷贝数的预测存在背景噪音。将长度>600bp 的片段中的 SNP 排除后，除了为二倍体的假常染色体区域（PAR）外，得到其余区域的拷贝数=1。当相同的样本与 FFPE 参照进行比较时（右列），拷贝数片段大小比较图提示不存在片段长度的偏差，并且表明所有的 SNP 都可以被纳入进行拷贝数的检测，但是 HMM 图和拷贝数预测仍然存在背景噪音

## 结论

当使用 Mapping 芯片用于 FFPE 的 DNA 样本以识别拷贝数的改变时，DNA 的降解和污染都是特别的难题，但是这些依靠计算机都可以很大程度的进行克服。在计算机上进行的步骤包括对于片段大小偏差的补偿，以及对较大片段中 SNP 的排除 CNAT4.0 和 CNAG 都可以自动进行片段大小偏差的校正（免费下载的拷贝数分析软件工具）。使用 CNAG 可以进行片段大小的切点值的选择，相应的排除设定可以在任一上述工具中进行。虽然对于 FFPE 的样本来说，由于任何片段大小的排除造成基因组的覆盖情况的减少，但是对于分析中保留的 SNP 仍可进行可靠的拷贝数的预测。因此，这些修正能够对 FFPE 样本进行全基因组、高分辨率的拷贝数检测。

## 参考文献

Bignell G.R., Huang J., Greshock J., Watt S., Butler A., West S., Grigorova M., Jones K.W., Wei W., et al. 2004. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14:** 287–295.

Huang J. 2004. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics* **1:** 287–299.

Huang J., Wei W., Chen J., Zhang J., Liu G., Di X., Mei R., Ishikawa S., Aburatani H., Jones K.W., and Shapero M.H. 2006. CARAT: A novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics* **7:** 83.

Ishikawa S., Komura D., Tsuji S., Nishimura K., Yamamoto S., Panda B., Huang J., Fukayama M., Jones K.W., and Aburatani H. 2005. Allelic dosage analysis with genotyping microarrays. *Biochem. Biophys. Res. Commun.* **333:** 1309–1314.

Jacobs S., Thompson E.R., Nannya Y., Yamamoto G., Pillai R., Ogawa S., Bailey D.K., and Campbell I.G. 2007. Genome-wide, high resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using micro arrays. *Cancer Res.* **67:** 2544–2551.

Kennedy G.C., Matsuzaki H., Dong S., Liu W.M., Huang J., Liu G., Su X., Cao M., Chen W., Zhang J., et al. 2003. Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21:** 1233–1237.

Komura D., Shen F., Ishikawa S., Fitch K.R., Chen W., Zhang J., Liu G., Ihara S., Nakamura H., Hurles M.E., et al. 2006. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* **16:** 1575–1584.

LaFramboise T., Weir B.A., Zhao X., Beroukhim R., Li C., Harrington D., Sellers W.R., and Meyerson M. 2005. Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput. Biol.* **1:** e65.

Lin M., Wei L.J., Sellers W.R., Lieberfarb M., Wong W.H., and Li C. 2004. dChipSNP: Significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20:** 1233–1240.

Nannya Y., Sanada M., Nakazaki K., Hosoya N., Wang L., Hangaishi A., Kurokawa M., Chiba S., Bailey D.K., Kennedy G.C., and Ogawa S. 2005. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.* **65:** 6071–6079.

# 23　遗传关联研究中显著性的评估

Mark J. Daly

*Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114; Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142*



## 简介

估计显著性差异是每一个遗传相关研究至关重要的一步，通常是以 $P$ 值的大小来表示显著性差异，大多数研究基于 $P$ 值的大小对遗传变异的检验得出一个定性的结论，即遗传变异与疾病之间是否存在相关。本章主要介绍显著性检验的基本概念，以及设计恰当的显著性检验的基本策略，并进一步讨论显著性检验能否得出遗传相关的结论。

### 基本的统计概念和统计方法

显著性检验或者假设检验是数学概念中应用最为普遍之一。一般来说，当要进行一项实验研究，就要运用假设检验。通过实验得到的观测数据，计算出假设成立的可能性。在计算假设成立可能性的过程中，可能有多种多参数的数理模型，那么如何选择合适的数理模型，就需要进行可能性的比较。在实验观测数据的基础上，计算每个模型的可能性，并根据不同模型的复杂性的比较，最终确定一个最符合研究对象数据的数理模型。在遗传相关性研究中，常见一个普遍性问题，研究的基因 X 是与疾病相关或与研究感兴趣的基因表型相关呢？

虽然这一问题抓住了大多数相关性研究的核心，但是在许多重要的方面并不准确。①许多研究不仅仅关注某一个基因，而是全基因组（如全基因组相关性研究）和某种研究的疾病，这些疾病是与少数几个基因相关还是与多数基因相关至今还不清楚。②在相关性研究中，基因本身不能直接被检测，但这些基因实质上有功能。遗传变异对一个基

因产物的影响微乎其微（如可辨别的蛋白质编码的改变或者不可变剪接点的改变）。大量的多态性即使是邻近所研究基因的功能编码区，也不能解释这些多态性位点对基因的影响。这就导致研究问题的复杂性，如何构造假设检验去论述多种变异位点以及他们的联合作用；如何解释由多重假设检验得出的结果。③即使所研究的遗传变异点只有一个，同样会有一个问题，没有一种特定的数理模型能解释这个基因或者变异点与基因表型的关联。

基因或变异点与基因表型关联的研究，虽然还没有特定的数理模型能直接对其进行分析，但有一种方法能解决这个问题。例如，研究某 SNP 与某个疾病是否存在关联，该 SNP 是基因编码区的变异（如 SNP 位点 R、W 的变异）。我们收集了 1000 例患该病患者以及 1000 例对照健康人的 R、W 基因型数据，然后通过比较 R、W 基因位点在病例组与对照组频率进行分析。尽管没有特定的模型分析这种相关性的问题，但可构造一个模型替代该 SNP 与疾病不相关的零假设即 SNP 位点 R、W 的变异在病例组和对照组的频率是相同的。这样就把问题转化成了在病例组与对照组 R、W 的变异频率相同时是否拒绝零假设的问题。不是计算 SNP 与疾病相关的确切值，而是运用最基本的显著性检验方法，可以计算出假设基因型频率在两组人群相同的情况概率值的大小。

|  | 基因型 W | 基因型 R |  |  |
|---|---|---|---|---|
| 病例 | 246 | 1754 | 2000 | 12.3% |
| 对照 | 200 | 1800 | 2000 | 10.0% |
|  | 446 | 3554 | 4000 |  |

$\chi^2$＝所有表格（O－E)$^2$/E 的总和

O 是每个表格的观察值、E 是每个表格的期望值

自由度≥（行－1）（列－1）＝1

$P$（$\chi^2$）≥5.34，$df$＝1）＝0.02

对于这个问题，我们通常应用 Karl Pearson 卡方拟合优度检验，通过对观测值与期望值进行比较，得到一个在已知分布中的偏离值即 $\chi^2$ 值，根据 $\chi^2$ 值就能得出一个经典概率值 $P$。许多参考书中对这一基本统计方法有更详细的说明（Freedman et al. 1998; Venables and Ripley 2002)。2001 年 salsburg 著的《女士品茶》中完整展现了 20 世纪统计发展史。卡方检验方法提出不久，R. A. Fisher 就提出了许多显著性检验基本概念，包括概率值 $P$ 值这一重要概念，即当概率值 $P$ 足够小时就可以拒绝零假设。根据小概率事件在一次实验中不可能发生的原理，在统计学上通常是以 $p$ 值大于或小于 0.05 作为拒绝或接受零假设的标准，0.05 也是发生一类错误的概率值。在上例中，$P$ 值为 0.02，可以认为在 50 次实验中，只有 1 次出现例子中观测的结果，那么则发生了小概率事件或者是偶然事件，那么认为病例与对照中基因频率相同的假设是成立的可能性很小，则病例与对照组中基因频率存在差异。在相关性检验中 2×2 四格表的卡方检验最简单最常见。Fisher 检验（适用于单元格中有一个或多个期望值小于 5)、Cochran-Mantel-Haenszel 检验（根据某些变量如不同的人群将研究资料分成多个 2×2 四格表的进行卡方检验）(Agresti 1990) 等方法也常用在相关性研究中，结果的解释与 2×2 四

格表的卡方检验相似。

在上例中，尽管 $P$ 值的计算并不难，但要得出一个定性的结果相当的复杂。一方面，因为研究的目的是建立遗传变异与疾病表型间的联系，并将其作为生物学进一步研究和疾病防治工作的基础。研究的目的并不仅仅局限于基因组本身，也不是单凭某个统计阈值去拒绝零假设。试图验证已报道的发现时，研究人员应清楚当拒绝不相关的模型时什么样的假设是恰当的，那些希望探索医学细节性知识的研究人员更应该清楚这一点。按严格的数学术语说法，Neyman 和 Pearson 指出研究人员在零假设和备择假设中只能接受或拒绝一种假设，而不能不比较两者就凭空地接受或拒绝其中的一种假设。许多观点来源于这种思维并且这种思维影响下面讨论的问题。

在早期研究相关问题，还有一些较为常见的数学模型，如 Logistic 回归模型作为一种正式的假设检验方法。这里建议一种单参数的数理模型（模型中参数是表示由于相关的等位基因拷贝数增加而引起疾病危险度增加的对数值），并将有单参数或无单参数（利用标准化的数学算法获得这个参数最可能的估计值）的模型与观测的结果（如病例和对照组的情况）相比较得出两者的相似度，相似度的差异可以通过卡方检验的方法获得，即

$$2ln\text{（有合适参数模型相似度/无单参数模型相似度）}$$
$$=\chi^2\text{（}N\text{：自由度，}N\text{＝多余的参数数量）}$$

这样就可以用具体的假设检验方法重新阐述这个问题，虽然这种方法还没有进一步的具体化，但是相关研究的一种突破。在今后研究中，具体的假设检验或多或少会应用到。上面提到的回归模型更直观，因此广泛地应用于定量表型的分析，线性参数用以来描述研究中不同等位基因对表型差异性的影响，在此基础上进而分析模型的相似度及显著性。通常在模型中除主要的参数外会增加一些参数，这样会增加模型分析的自由度，会加大统计分析的量，但是能够调整模型，使其更合理。

## 遗传关联研究中的特殊问题

本节并不是阐述遗传关联研究中有关显著性估计的一些基础问题。当研究人员考虑单个 SNP 或者其他的遗传变异与疾病相关的问题时，如果仅单个 SNP 与疾病相关，首先，我们不能认为这个变异点或者与变异点连锁不平衡的点同疾病间存在真正的相关。因为人类基因组中有 2 万多个基因和上百万个遗传变异，其中只有极少个别基因或者遗传变异与某种疾病关联。因此，当研究人员真正试图分析遗传变异与某疾病的相关或者不相关时，收集数据前，一定要慎重考虑接受备择假设的可能性。例如，一个研究者或者一个研究团队同时研究多个基因或者数以万计 SNP 的全基因组相关研究时，不同于单个假设检验，每个基因中的多个遗传变异位点都要估计，因此接受备择假设的统计阈值更为严格，一般 $P$ 值更小才能不拒绝备择假设。有关全基因组相关研究中统计分析的注意事项在 Hirschhorn 和 Daly（2005）以及 Wang 等（2005）出版的论著中均有介绍。

正是由于这些必须考虑的因素，才有了大家经常讨论的“多重检验”问题。近些年来，“多重检验”带来了相当多的方法学上的考虑。简而言之，在无效假设前提下，最

好的检验结果就是能得出一个极显著的 $P$ 值（不足为奇，在 $N$ 个相互独立的实验中得出的最佳 $P$-值为期望 $P$ 值的 $1/N$）。因此，在解释一个观测结果的显著性意义时，必须考虑一些问题，尽管多重检验并不能很敏锐地应用于生物学的所有领域中，遗传学研究人员在进行全基组扫描（最初目的是进行连锁分析，现在目的是进行相关研究）时，已经充分认识到了多重检验问题。这也可能是由于全基因组扫描研究不再需要进行预先假设，而在以往的临床试验研究或流行病学研究中，环境影响评估中必须要有无效假设。

## 置换检验

习惯上有显著性差异意义的统计阈值概率值定为 0.05，全基因组相关分析时为了得到显著性差异的结论，统计阈值概率值应定义为 $0.05/N$，一种简单的方法就是运用 Bonferroni 方法进行校正。为了达到相同的结论，$N$ 次检验研究人员可能会成倍的地增加了 $P$ 值（实际上，$0.05/N$ 只是假设检验时一类错误 $\alpha$ 概率值的大概值，其中 $\alpha$ 满足 $0.05=1-(1-\alpha)^N$），同时，由于在 $N$ 次检验时的自由度没有考虑，这样校正因子在遗传相关研究中可能会使研究结果过保守。SNP 可能与另外的一些 SNP 相关（在连锁不平衡传递模式中），这将增加期望得出的最小 $P$ 值。而使用置换检验的方法得出的全基因组相关分析的显著性差异的统计阈值更加准确，置换检验的方法要求研究的数据集完整，每个数据都匹配所有的属性，即研究中每个观察对象随机转换得到的表型值，如病例对照的标签（还包括标记物间的 LD 值、散在的数据误差及缺失值），进而反复多次地统计分析由原始数据置换数据。用这种方法分析大量的置换数据的研究，提供了极限值准确的统计阈值。

这类检验方法主要是控制实验假设检验犯一类错误的概率值，即阳性率概率值。某个基因与某疾病存在潜在关系时，就迅速推出复杂研究中疾病病因学似乎有道理的功能假说，这存在一些问题。然而，有显著性意义的统计阈值尽管是作为解释统计结论的有力证据，但是在全基因组关联的研究中统计阈值可能过低。假设我们检验 500 000 个 SNP 时，可能推荐使用 $10^{-7}$（$=0.05/500\ 000$）的概率值作为有显著性意义的统计阈值，那么在单个研究中，统计阈值极大地限制了相关证据的检验功效。如一项有 1000 病例、1000 正常对照的病例对照研究，若在 $p<0.01$、等位基因的相对危险度为 1.5 的条件下，能有 95%的概率检出 10%的等位基因；同样的研究，若在 $p$ 小于 $10^{-7}$ 时，仅有 12%的概率检出相同比例的等位基因，明显检验的效能降低。在这种情况下，通常给观测结果显著性差异的统计阈值一个折中范围，如 $0.00001<p<0.001$。在一般的研究中，$0.00001<p<0.001$ 时表明观测结果有显著性差异，但是在 500 000 次检验之后，会随机产生成千上万个相似的结果满足 $0.00001<p<0.001$。这有两个原因，①$p$ 值满足统计阈值概率值的范围的单个研究的结果并不能确定是否是随机产生的，因此 $p$ 值有意义并不等同于存在真正的相关性。②单个结果或多个结果的 $p$ 值在这个阈值概率范围内，但对整个 $p$ 值的分布影响不大。

## 错误发现率分析方法

错误发现率分析方法（FDR）是校正多重比较的另一种方法，取代控制阳性率概

率的方法（如 Bonferroni）。这种方法由 Benjamini 和 Hochberg 两位研究者于 1995 年首次定义，现在 FDR 方法有所改进。这种方法不同于 Bonferroni 校正和置换检验的方法计算每个检验的 $P$ 值，而是评价假设存在多个中量阳性信号的 $P$ 值的分布。相对于我们开始讨论的显著性检验方法即评价假设不存在相关的显著性的概率值，FDR 方法是在假定的显著性水平下直接的估计存在相关性的概率值。具体化这一问题，假如前面 500 000 个 SNP 的研究中，没有一个 SNP 满足显著性差异的阈值，即 $P$ 值大于 $10^{-7}$，但其中有 10 个 SNP 的 $P$ 值为 $10^{-6}<P<10^{-7}$。根据 $P$ 值空分布，至少有一个检验随机小于 $10^{-6}$，尽管其中没有一个小于 $10^{-7}$，但前 10 个 SNP 中的大部分点构成真正的阳性。

如何评价这一结果，置换检验的重要性不可忽视。标记点之间的 LD 相互依赖、数据质量不可避免的缺陷、人群结构和不相关样本间的远亲关系能改变分析结果的期望值。尤其是全基因组的相关研究，研究人员应该确定分析的结果是否与 $P$ 值空分布保持高度一致，而不是推测其结果的一致性（大多数的分布应与 $P$ 值空分布一致，因为在全基因组研究中可能有极少的阳性相关）。

FDR 和相关的检验试图估计和控制假阳性的比例，而不是消除假阳性。在上面的例子中，FDR 分析得出这样一个结论，如果认为 $P$ 值最小的 10 个 SNP 有意义，那么这个推断可能有 5%～10%的概率是不成立的，即出现假阳性。尽管研究者可能不希望进一步探讨在单个独立研究中还没确证的相关，但是在研究中清楚显示多个基因因子的强有力证据，毫无疑问在以后的研究中很快能得到肯定的结果。更具体一点说，研究者设定一个可以接受的错误发现率，FDR 步骤将得到满足可接受错误发现率标准的最大的一组结果。2003 年 Storey 和 Tibshirani 外延了这一概念，并提出了现在广泛应用的 $q$ 值。例如，用 $q$ 值来定义全基因组数据集的每个 SNP 结果，$q$ 值表示假阳性比的预期值的平均值，特别是这一结果被用着为阈值。在上面的例子中，第 10 个 SNP 的 $q$ 值为 0.1，我们可以正式地说前 10 个 SNP 与检测表型相关的可能性的概率值为 90%。FDR 这种方法被广泛、高效地应用于基因芯片分析，分析比较两个可能有成千上万个基因变异对象。实际上与表型真正相关的数量可能相对较小，因此从空分布不能推导出大量的偏离。然而，在解释全基因分布最显著性双尾方面这些方法证明似乎有用。

## 其他方法

其他的一些方法补充了那些解决相关和多重检验问题，鉴于不断涌现的海量数据许多学者推崇调整每个独立检验的先验概率（基于先前的数据、效能、LD 模型）的方法进行全基因组研究（Wacholder 2005、Pe'er et al. 2006）。在候选基因研究中，无论其是 SNP 基础上或是单倍型基础上的研究，将基因的一系列检验作为一个单一的组进行相关研究和上位效应检测（当然，随着研究的不同会有差异）。许多学者指出 Hotelling's $T^2$ 统计是一个强有力的统计方法分析多个标记（Xiong et al. 2002），尤其是在与表型相关的单个基因的多个相互独立的等位基因的研究。这种方法还没有考虑到用于基于独立的标记点或者单倍体型的检验。Patterson 建议全基因组的贝叶斯逻辑分析，依照这种方法完整的实验能探索出基因组中任何位置存在相关的证据（Patterson

et al. 2004)。依照不同相关研究的本质、及研究的不同范围，不同的假设条件，能得到不同的有用先验信息等情况，以上的每一种检验方法都可能有用。

## 结论

本章主要阐述遗传相关研究的背景知识以及在具体的实验中如何设计并估计显著性检验的参考方法，并没有全面的介绍方法学及阈值的应用问题。而是紧紧围绕显著性这一基本概念，因为良好的理解有助于遗传学者更好的研究。尽管外理数据的软件和方法能返回一个 $P$ 值，但更重要的是要理解怎样提出一个问题，然后怎样去解释研究的结果，达到发现有效的生物学相关的目的。

## 致谢

作者在写此部分内容时，经常与 Purcell 和 Nick Patterson 进行细致深入的讨论，并且得到了评价和指正，在此表示感谢。同时还要感谢 Clay Stephens 对本文的评阅和指正。

## 参考文献

Agresti A. 1990. *Categorical data analysis.* John Wiley and Sons, New York, pp. 100–102.

Benjamini Y. and Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57:** 289–300.

Freedman D., Pisani R., and Purves R. 1998. *Statistics*, 3rd edition. W.W. Norton, New York.

Hirschhorn J.N. and Daly M.J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6:** 95–108.

Patterson N., Hattangadi N., Lane B., Lohmueller K.E., Hafler D.A., Oksenberg J.R., Hauser S.L., Smith M.W., O'Brien S.J., Altshuler D., et al. 2004. Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74:** 979–1000.

Pe'er I., de Bakker P.I.W., Maller J., Yelensky R., Altshuler D., and Daly M.J. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.* **38:** 663–667.

Salsburg D. 2001. *The lady tasting tea: How statistics revolutionized science in the twentieth century.* Henry Holt and Co., New York.

Storey J.D. and Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100:** 9440–9445.

Venables W.N. and Ripley B.D. 2002. *Modern applied statistics with S*, 4th edition. Springer, New York.

Wacholder S. 2005. Publication environment and broad investigation of the genome. *Cancer Epidemiol. Biomarkers Prev.* **14:** 1361.

Wang W.Y., Barratt B.J., Clayton D.G., and Todd J.A. 2005. Genome-wide association studies: Theoretical and practical concerns. *Nat. Rev. Genet.* **6:** 109–118.

Xiong M., Zhao J., and Boerwinkle E. 2002. Generalized $T^2$ test for genome association studies. *Am. J. Hum. Genet.* **70:** 1257–1268.

# 24 评估人类变异数据以探索自然选择标记

Mike Bamshad[1] and J. Claiborne Stephens[2]

[1]*Departments of Pediatrics and Genome Sciences, Division of Genetics and Developmental Medicine, University of Washington School of Medicine, Seattle, Washington 98195;* [2]*Motif BioSciences Inc., New York, New York 10017*



## 简介

在许多方面，人与人之间存在差异，如人的相貌、行为以及对疾病的易感性等千差万别。这些表型的差异，一部分来源于环境因素对机体的影响，另一部分是由于个体遗传背景的差异。不同的遗传变异一起构成了人类基因组中的一部分，而正是这一部分因自然选择而表达不同。对机体有益的基因变异可能经过一代又一代的复制和传递，基因变异的频率越来越高，这种基因的自然选择称为正选择或适应性选择；反之，对机体有害的（无论对机体的危害是中度还是严重的）基因变异在进化的过程中则逐渐减少，这种自然选择称为纯化选择。

适应性的自然选择最先由 Darwin 和 Wallace 提，并认为在过去的 100 000 年人类进化的过程中自然选择对人类变异起着尤为重要的作用。在这期间，解剖学意义上的现代人类从非洲迁徙，并适应了不同的自然环境（Darwin and Wallace 1858; Klein 1999)。现存的变异以及在进化过程中将不断出现新的功能性变异，将更利于人类在自然环境的生活，这些功能性变异就是适应性自然选择的潜在目标。一个重要的推论：适应性的自然选择似乎是一种神秘的魔法棒，使在对表型几乎没有影响的变异的遗传变异的背景下（如中性变异），它能从遗传变异中选出功能性的变异，从而，易于选择的变异给我们提供了一些启示，什么样的基因和什么样的无编码调节功能成分影响人类表型变异，包括个体健康的差异。因此，理解进化力量和人口统计过程如何形成了人类遗传变异的地理分布，为解释遗传变异的模式提供了方向并为试验策略鉴定疾病相关的变异的奠定了理论基础。

发现自然选择在人类基因中的作用的最重要信息来源之一是基因的本身，人口遗传模型预测不同类型的自然选择对同一个基因变异将产生不同的信号。因此，由变异的模式能推导出潜在的自然选择作用。近年来因人类及其他物种的海量的新变异数据信息可用，这一策略得到广泛应用。适应性选择或者平衡选择对基因是否产生作用，这一推论的有效性已在人工模拟自然选择条件下的现代的生物组织和国内的动植物（如棉花）实验中，经遗传变异模式的统计分析得到了确证（Doebley et al. 2006）。很显然，这些实验不能在人类身上进行实验，但是也没有理由去怀疑使用相同的统计方法分析人类变异数据得出推论的有效性。

本章，我们主要强调不同类型的自然选择中的一部分，及其对 DNA 变异模型的作用以及检验这些作用常用的统计检验方法。我们同时也介绍一些检测自然选择信号的策略，并对其检测的优点和缺点进行分析解释。这些策略可以应用于从实验数据到与疾病易感性存在差异的基因变异。其中的一些策略方法被推荐进行基因组扫描，以期得到自然选择的证据。最后我们将讨论与自然选择信号相关的一些问题，并讨论如何作出自然选择本质的推论、重要哲学问题（如何选择的单元）、理论性的概念以及其他物种的实验研究超出本章的介绍的范围，如读者有兴趣，想进一步了解的话，可以参考大量的相关资料（如 Sober 1993；Li 2006）。

## 鉴别选择的方法

基因突变是基因变异的主要来源，但是新的基因突变的结果是由自然选择和种群史决定的。选择结果使新的突变减少，或者保持一个相对适中的频率，或在群体中达到稳定（当突变的频率达到 100%时这一变异在群体中稳定传递）。当自然选择对遗传变异产生作用时，它改变 DNA 周围序列的变异模式，那么就会产生一个分子标记。这些分子标记可以有多种方法进行鉴别，并且有许多统计检验方法用来鉴定。在许多优秀的综述中对统计检验方法的应用有详细概述（Kreitman 2000；Schlotterer 2002；Vallender and Lahn 2004；Nielsen 2005；Biswas and Akey 2006；Harris and Meyer 2006；Sabeti et al. 2006）。

经典的统计检验方法分析自然选择是基于与期望值的比较，期望值是在概率模型的空分布（如参数方法）、基因组变异的经验分布或者两者的结合的条件下得到的。人口统计模型和参数的限制条件下模拟确定效能（Simonsen et al. 1995；Fu 1997；Wall 1999；Prezworski 2002；Sabeti et al. 2002）。这很重要，因为种群史，如亚人群、人口减少的时期（如瓶颈期）、人口的增长都会导致 DNA 序列变异的模式，而因这些原因引起的变异模式与因自然选择所引起的结果是相似的（Ptak and Przeworski 2002）。例如，低频率变异（如罕见的等位基因）的比例在人口激增的情况下预期会增大，因为新的突变正以较低的频率消失。实质上，人类这种罕见变异的增加通常会被解释为人口数量的激增所引起的。然而，正选择同样也会产生低频变异激增的相似的一个结果（Braverman et al. 1995）。因此，了解种群史是鉴定易于受自然选择影响的基因至关重要的环节。并且在所在的统计检验过程中，应考虑一些因素，如变异数据是如何

确定的以及当地的遗传因素的影响（如重组率）等。

## 比较物种间的多态性

通过比较人类与其他物种间的变异模式，我们已了解了一些进化对人类基因的作用。具体地说，在人类漫长的进化过程中，适应性选择增加了对人类有益的功能变异的递率。这一增加可以通过比较系谱中的非同义（如氨基酸的改变）或者同义变化来鉴定（Bush 2001；Yang 2002；Wong et al. 2004）。目前大量的 DNA 序列数据资源共享，这种方法也日益流行，尤为强调的是，通过比较人与猩猩不同进而得出适应性选择的证据（Bustamante et al . 2005；Nielse 2005）。不同位点的确定可能给进化本质的研究带来更深的启示，如语言的起源在现代人类进化过程中的适应起着重要的作用（Enard et al. 2002）。

$K_A/K_S$ 检验及 McDonald Kreitman 检验的统计方法用来鉴别标记物（Li et al. 1985；Nei and Gojobori 1986；McDonald Kreitman 1991）。一般认为，这些检验方法较为保守，因为代替率是由已检测的所有氨基酸位点归纳得到的，并且这些氨基酸位点可能受不同功能限制。相似的检验也应用于分离蛋白质重要功能域、氨基酸残基和非编码调节序列（Suzuki and Gojobori 1999）。

## 比较物种内的多态性

大量的功能变异明确了基因组中自然选择针对某一物种作为一个整体的一段区域。这类标记提供了重要的信息：表明这种物种如何进化从而区别于其他的物种。然而很少能够揭露自然选择在同一个物种的个体间的表型差异（如外表 疾病的易感性）形成过程中是否起到作用。为了研究自然选择是否影响对个体内在差异，比较同一物种的不同群体的多态性数据来寻找自然选择的标记（Bamshad and Wooding 2003）。研究新的统计检验的方法来鉴定这种标记，目前是一个非常热的研究领域，在过去的几十年中，出现了许多不同的统计方法。这些统计方法广义上根据探索的目的分为几类：①遗传多样性的增加或减少 ；②群体间不存在变异或者大量变异；③扩增的单倍型。

除自然选择外还有两种不同的选择模型，标记显示这些基因附近区域的遗传多样性明显下降。第一种模型即背景选择（background selection），清除对机体有害的突变并减少附近区域的变异（Charlesworth et al. 1993），而背景选择的作用因基因重组率、选择的强度和基因的突变率变化而变化（Hudson and Kaplan 1995）。第二种模型即选择牵连效应（genetic hitchhiking），当群体中发生对选择有利的基因突变时，与之紧密连锁的位点也会发生相应的基因频率变化，结果有利等位基因的频率上升，不利的等位基因频率下降（Maynard Smith and Haigh 1974；Fay and Wu 2000），也就是所谓的选择性清除现象（selective sweep）。遗传的多样性最终会恢复，但是需要一个漫长的过程，因为新基因的突变率很低。因此与基因组其他区域相比连锁区域的大量罕见的变异在适应性的自然选择的作用下导致遗传多样性的整体下降。发现这类标记常用的统计方法有 Tajima'D（Tajima 1989）、Fu 和 Li's $D^*$（Fu and Li 1993）及 Hudson-Kreitman-Aguade（Hudson et al. 1987）。

例如，药物代谢酶常见的编码基因 *CYP3 A4* 和 *CYP3 A5* 附近的基因区域，其他位置相比出现了大量的罕见等位基因，表明了遗传多样性的下降（Thompson et. 2004）。自然选择具体的靶点并不清楚，但是影响盐水内环境平衡和高血压的 *CYP3 A5* 功能性等位基因频率出现独特的地理分布，基因的频率与距离赤道的物理位置相关。不连锁与高血压相关基因 *AGT* 的一个变异也出现相似的关系。这一结果表明与纬度相关的选择压力对这两个基因均有作用。基因组中其他一些区域也缺乏多样性，其中有黑皮质素-1 受体基因 *MC*1*R*（melanocortin-1 receptor；Makova et al. 2001）、人乳糖酶根皮苷水解酶基因 *LCT*（lactase phlorizin hydrolase；Bersaglieri et al. 2004）和 *Kel* 抗原基因 *KEL* 等（Kel antigen；Akey et al. 2002）。

在选择清除中，因选择牵连效应，一些与对机体有益的等位基因连锁的新突变频率上升，但不能达到稳定，即当突变的频率达到 100％时这一变异在群体中稳定传递。这些新等位基因相对已存在的等位基因出现得比较晚，可以通过与已存的等位基因比较可以对新等位基因进行预测，因此另外的一些正选择标记存在于含有大量高频突变等位基因的区域（Watterson and Guess 1977）。基于与另一个群体或物种进行比较，可以推断出等位基因是遗传而来的还是突变而来的。黑猩猩和大猩猩是与人类最近的两个物种，有着共同的祖先，因此这两个物种用来作为对照最为普遍。常用的确定含有大量高频突变等位基因区域的检测方法是 Fay 和 Wu 的 H 法（Fay and Wu 2000）。

例如，编码药物代谢酶基因 *CYP1A2* 的 4 个 SNP，在黑猩猩和大猩猩两物种中是固定的等位基因（Wooding et al. 2002）。在人类这四个常见的 SNP 每个频率均大于 90％，因此，认为这几个 SNP 是突变而来的。通过中性模型估计在每一个频率下变异点的期望分数，并进行比较，显示 *CYP1A2* 基因有大量的高频或低频的突变等位基因，这说明 *CYP1A2* 基因可能既受到正选择又受到近期人口增长的影响，尽管这两个因素相对强度并不清楚。

受清除影响基因组区域的大小由选择的强度和当地基因的重组率所决定的。强大的选择优势导致的基因清除或者低重组率的区域，可以影响基因组中较大的区域，一方面为选择标记的发现提供了方便，另一方面也增加了发现随机变异的难度。正如前面所提到的，人口规模的激增也能产生大量罕见的等位基因，这些等位基因的产生是正选择的作用还是人口历史的原因很难区分。

当一个突变点出现时，在现存的背景下单倍体型满足新突变点与连锁的多态性位点间完全连锁不平衡（LD）。随着时间的推移，新突变和基因重组逐渐减少了这类单倍体。一般来说，旧的突变点的和典型普通的突变点在更小的单倍型中（如在突变和连锁的多态性间仅存在短片段的 LD 区域）。新突变点或者低频突变点可能与小的或者大的单倍型相关。但是，由选择清除产生的等位基因既有较高的频率，又可能出在较大的 LD 区域。

选择标记的统计检测方法已成为了研究的热点（Sabeti et al. 2002；Kim and Nielsen 2004；Hanchared et al. 2006；Voight et al. 2006；Wang et al. 2006）。其中最流行的统计检测方法是用于长片段的单倍型检测（LRH）。该统计方法能使正选择作用的基因组区域定位到一个相对较小的区域。该统计方法有一个限制条件就是只能检测时间相对相近的选择标记，然而在人类由原始人转变为现代人的 100 000 年中，当前的

选择性进化尤为重要，因此这一方法还是广为应用。

例如，断奶后大多数哺乳动物包括人类失去了代谢乳糖的能力。因此人类乳糖酶的非持续性似乎是遗传的，并且在世界上大部分地区非常典型（Swallow 2003）。然而，在北欧许多种群中乳糖酶普遍持续表达（如瑞典和丹麦人群中超过 90%）；撒哈拉以南的非洲地区的人群也出现相同的情况（如图兹人群也超过 90%）；这些地方奶制品是主要的食物来源。长久以来人们怀疑 LCT 是正选择的目标基因，恰好正选择与奶牛的驯养相一致（Hollox et al. 2001；Bersaglieri et al. 2004；Myles et al. 2005）。在欧洲人群中，成人 *LCT* 基因表达存在差异，由 LCT 基因序列上游 14kb 处-13910 位置的 SNP 的 C/T 等位基因变异引起（Enattach et al. 2002）。而撒哈拉以南的非洲地区人群存在三个新的 SNP，位于 *LCT* 基因序列上游，分别是 G/C-14010、T/G-13915 和 C/G-13907，在体外实验中似乎增加了 *LCT* 基因的表达（Tishkoff et al. 2006）。这些变异点与欧洲和非洲人群延长的连锁不平衡区域相关，表明了近期正选择在至少世界上有两个地方的人群中快速增加了 *LCT* 基因几个不同变异点的频率。每一种情况下，选择压力似乎表现对更高的适应度有一种适应性反应即提供消化奶制品的能力。

当地适应进化可能使一个群体有利的等位基因的保持高频率，但在另一个群体中频率不一定升高。群体间的等位基因频率存在较大的差异，这一差异是正选择的另一种潜在标记。例如，对间日疟原虫有抑制作用的趋化因子受体的等位基因 *FY** *O*，在撒哈拉以南的非洲地区人群中几乎稳定传递，而在非非洲人群中则很罕见（Hamblin et al. 2002）。有意思的是，*FY** *O* 等位基因单独出现并在新高加索高地和撒哈拉以南的非洲地区中易于受选择的影响，这种会聚性的进化可能是自然选择作用最强有力的证据。苦味感受器（Soranzo et al. 2005）基因 *T2 R16* 、编码人类色素沉着起作用的离子交换器的基因 *SLC24 A*，两者毗邻的基因组区域在非洲和非洲以外的人群中存在差异。这种差异仅出现在这些人群，他们至少是隔离群体繁殖后代的。因此人类的差异与自然选择相关，因解剖意义上的现代人类，100 000～50 000 年前由非洲迁移出来。

在一个或多个群体中，自然选择在一个位点上倾向保存两个或多个等位基因。这就是所谓的平衡选择，等位基因的频率由于某些罕见等位基因优势的原因保持平衡。实质上，平衡选择使群体间产生比预期很少的差异，并产生了大量中等频率的等位基因，因为选择牵连效应使连锁位点的变异累积（Lewontin and Hubby 1996；Kaplan et al. 1988）。在许多动植物中，平衡选择似乎对保持一些位点的多样性起着一个重要的作用，这些位点主要协调识别自身或非自身的组织（Richman and Kohn 1999）。在人类这些位点已有较透彻的研究，如宿主病原体反应的位点，主要包括人类白细胞抗原 *HLA*-Ⅰ 和 *HLA*-Ⅱ基因（Hughes and Yeager 1998）、*TAS 2 FR38* 基因（Wooding et al. 2004）、*G6PD* 基因（Verrelli et al. 2002）和 *CCR5* 顺式调控区域等（Bamshad et al. 2002）。

## 基因组扫描技术

用来检测单个位点选择作用的方法也可应用于全基因组扫描发现选择的标记（Lewontin and Krakauer 1973）。在过去的几年中，国际人类基因组单体型图计划（Inter-

national HapMap Consortium 2005)、Perlegen Sciences 计划（Hinds et al. 2005)、Applera 计划（Bustamante et al. 2005）和 Genaissance 测序计划（Stephens et al. 2001）提供了海量可用的基因多态性数据。目前，将近十几种选择作用标记的全基因组扫描已完成（Cargill et al. 1999；Sunyaev et al. 2000；Akey et al. 2002；Payseur et al. 2002；Bamshad and Wooding 2003；Bustamante et al. 2005；Carlson et al. 2005；Weir et al. 2005；Bubb et al. 2006；Voight et al. 2006；Wang et al. 2006)。它们的研究策略各异，例如，使用不同的标记物、不同的统计方法和或扫描相对编码区隐匿的基因组区域（McVean and Spencer 2006)。通过比较基因扫描全面的信息，给自然选择对基因组的整体的作用加以归纳总结，并提供迄今仍未知的功能变异基因信息。

已完成的人类基因扫描中大多数发现选择作用标记，基于群体间不同的等位基因频率来探索选择标记的许多基因扫描也已完成。例如，单倍型数据分析揭示了 27 个基因中 926 个 SNP 在群体间的差异比 *FY* 位点还明显（International HapMap Consortium 2005)，预计有 30 多个 SNP 造成非同义氨基酸替换，其中包括前面提到的 *SLC24A5* 基因中的 rs1426654，部分解释了非洲和非洲外其他人群间皮肤色素变异（Lamason et al. 2005)。Weir 等（2005）估计人群差异，用 HapMap 和 Perlegen 数据，发现大约 300 个含有选择作用的目标基因区域。许多其他类型的扫描方法已得到应用，与基于 LD 测度远程衰减原理的长片段单倍型检验相似的统计方法进行分析。Wang 等（2006）在 Perlegen 数据中发现聚集在 1799 个基因的 25 386 个 SNP，因这些 SNP 的 LD 远程衰减与全基因组平均水平有显著性的差别。Voight 等（2006）提出了一种相近的统计方法，并发现了 300 个区域，含有 455 个基因，估计这些基因为正选择的目标基因。

基因组扫描方法总共发现了 2300 个基因，怀疑它们是人类正选择的目标基因，其中的大多数先前并没有作为候选基因。实际上，先前报道的正选择的候选基因在基因组扫描中得到确证的仅不到一半。只有少数几个基因报道在多个分析中有选择标记，而且选择标记多数仅限定在单个群体中而不是多个群体间。其中一些目标基因可能是假阳性，因为各个实验设计存在差异。多个研究间直接比较有困难，有意思的是，许多基因组区域似乎是选择的目标，但它们几乎不含有基因。这个结果强调适应进化对非编码序列比先前认同适应进化在形成人类变异模式中，起着更大的作用（Ponting and Lunter 2006)。

## 结论

选择如何对人类基因组产生作用，我们的解释相对简单，实际上，选择的作用可能很复杂。选择的强度会因时间而波动，并且不同的选择作用间会相互作用；还有，群体遗传模型的预测是基于人口参数的估计的前提条件下进行的，然而人口参数是模糊的。因此，进一步明确选择作用的基因，需要清楚群体的人口结构。

对于已发现的大多数标记，进一步验证和确认其结果提出了新的挑战。例如，结果的重复很难，因为从相同的群体中单独抽样，含有重叠的祖先；不同的遗传变异方式并不是

相互独立的。证明一个选择标记是最客观的证据就是选择的目标基因有能影响表型的功能序列，并且认为表型是因自然选择作用引起的。在这种条件下，可以通过基于群体的相关研究、体外分子生化研究和模式生物来评估功能作用，当然，这可能是最复杂的方法。

在某种程度上，即使对基因变异功能显著性未知的情况下，这些标记能够很好地预测具有重要生物医学功能的基因变异位点。为了这个目的，群体变异的研究将继续是研究者和临床医生关注的方向。

# 参考文献

Akey J.M., Zhang G., Zhang K., Jin L., and Shriver M.D. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12:** 1805–1814.

Bamshad M.J. and Wooding S.W. 2003. Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4:** 99–111.

Bamshad M.J., Mummidi S., Gonzalez E., Ahuja S.S., Dunn D.M., Watkins W.S., Wooding S., Stone A.C., Jorde L.B., Weiss R.B., and Ahuja S.K. 2002. A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. *Proc. Natl. Acad. Sci.* **99:** 10539–10544.

Bersaglieri T., Sabeti P.C., Patterson N., Vanderploeg T., Schaffner S.F., Drake J.A., Rhodes M., Reich D.E., and Hirschhorn J.N. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74:** 1111–1120.

Biswas S. and Akey J.M. 2006. Genomic insights into positive selection. *Trends Genet.* **22:** 437–446.

Braverman J.M., Hudson R.R., Kaplan N.L., Langley C.H., and Stephen W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140:** 783–796.

Bubb K.L., Bovee D., Buckley D., Haugen E., Kibukawa M., Paddock M., Palmieri A., Subramanian S., Zhou Y., Kaul R., et al. 2006. Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173:** 2165–2177.

Bush R.M. 2001. Predicting adaptive evolution. *Nat. Rev. Genet.* **2:** 387–392.

Bustamante C.D., Fledel-Alon A., Williamson S., Nielsen R., Hubisz M.T., Glanowski S., Tanenbaum D.M., White T.J., Sninsky J.J., Hernandez R.D., Civello D., et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153–1157.

Cargill M., Altshuler D., Ireland J., Sklar P., Ardlie K., Patil N., Shaw N., Lane C.R., Lim E.P., Kalyanaraman N., et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22:** 231–238.

Carlson C.S., Thomas D.J., Eberle M.A., Swanson J.E., Livingston R.J., Rieder M.J., and Nickerson D.A. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* **15:** 1553–1565.

Charlesworth B., Morgan M.T., and Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289–1303.

Darwin C. and Wallace A.R. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *J. Proc. Linn. Society Lon. Zool.* **3:** 46–50.

Doebley F., Gaut B.S., and Smith B.D. 2006. The molecular genetics of crop domestication. *Cell* **127:** 1309–1321.

Enard W., Przeworski M., Fisher S.E., Lai C.S., Wiebe V., Kitano T., Monaco A.P., and Paabo S. 2002. Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418:** 869–872.

Enattah N.S., Sahi T., Savilahti E., Terwilliger J.D., Peltonen L., and Jarvela I. 2002. Identification of the variant associated with adult-type hypolactasia. *Nat. Genet.* **30:** 233–237.

Fay J.C. and Wu C.I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155:** 1405–1413.

Fu Y.X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147:** 915–925.

Fu Y.X. and Li W.H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133:** 693–709.

Hamblin M.T., Thompson E.E., and Di Rienzo A. 2002. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* **70:** 369–383.

Hanchard N.A., Rockett K.A., Spencer C., Coop G., Pinder M., Jallow M., Kimber M., McVean G., Mott R., and Kwiatkowski D.P. 2006. Screening for recently selected alleles by analysis of human haplotype similarity. *Am. J. Hum. Genet.* **78:** 153–159.

Harris E.E. and Meyer D. 2006. The molecular signature of selection underlying human adaptations. *Yearb. Phys. Anthropol.* **49:** 89–130.

Hinds D.A., Stuve L.L., Nilsen G.B., Halperin E., Eskin E., Ballinger D.G., Frazer, K.A., and Cox D.R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:**1072–1079.

Hollox E.J., Poulter M., Zvarik M., Ferak V., Krause A., Jenkins T., Saha N., Kozlov A.I., and Swallow D.M. 2001. Lactase haplotype diversity in the Old World. *Am. J. Hum. Genet.* **68:** 160–172.

Hudson R.R. and Kaplan N.L. 1995. Deleterious background selection with recombination. *Genetics* **141:** 1605–1617.

Hudson R.R., Kreitman M., and Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153–159.

Hughes A.L. and Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annu. Rev. Genet.* **32:** 415–435.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Kaplan N.L., Darden T., and Hudson R.A. 1998. The coalescent process in models with selection. *Genetics* **120:** 819–829.

Kim Y. and Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160:** 765–777.

Klein R.G. 1999. *The human career: Human biological and cultural origins*, 2nd edition. University of Chicago Press, Chicago, Illinois.

Kreitman M. 2000. Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* **1:** 539–559.

Lamason R.L., Mohideen M.A., Mest J.R., Wong A.C., Norton H.L., Aros M.C., Jurynec M.J., Mao X., Humphreville V.R., Humbert J.E., et al. 2005. SCL24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310:** 1782–1786.

Lewontin R.C. and Hubby J.L. 1966. A molecular appraoch to the study of genetic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural

populations of *Drosophila pseudoobscura*. *Genetics* **54:** 595–609.
Lewontin R.C. and Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74:** 175–195.
Li W.H., Wu C.I., and Luo C.C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2:** 150–174.
Li W. 2006. *Molecular evolution*. Sinauer, Sunderland, Massachusetts.
Makova K.D., Ramsay M., Jenkins T., and Li W.H. 2001. Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. *Genetics* **158:** 1253–1268.
Maynard-Smith J. and Haigh J. 1974. The hitch-hiking effect of a favorable gene. *Genet. Res.* **23:** 23–35.
McDonald J.H. and Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.
McVean G. and Spencer C.C.A. 2006. Scanning the human genome for signals of selection. *Curr. Opin. Genet. Dev.* **16:** 624–629.
Myles S., Bouzekri N., Haverfield E., Cherkaoui, M. Dugoujon J.M., and Ward, R. 2005. Genetic evidence in support of a shared Eurasian-North African dairying origin. *Hum. Genet.* **117:** 34–42.
Nei M. and Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Mol. Biol. Evol.* **3:** 418–426.
Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86:** 641–647.
———. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39:** 197–218.
Payseur B.A., Cutter A.D., and Nachman M.W. 2002. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* **19:** 1143–1153.
Ponting C.P. and Lunter G. 2006. Signatures of adaptive evolution within human non-coding sequence. *Hum. Mol. Genet.* **15:** R170–R175.
Przeworski M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* **160:** 1179–1189.
Ptak S.E. and Przeworski M. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.* **18:** 559–563.
Richman A.D. and Kohn J.R. 1999. Self-incompatibility alleles from *Physalis*: Implications for historical inference from balanced genetic polymorphisms. *Proc. Natl. Acad. Sci.* **96:** 168–172.
Sabeti P.C., Schaffner S.F., Fry B., Lohmueller J., Varilly P., Shamovsky O., Palma A., Mikkelsen T.S., Altshuler D., and Lander E.S. 2006. Positive natural selection in the human lineage. *Science* **312:** 1614–1620.
Sabeti P.C., Reich D.E., Higgins J.M., Levine H.Z., Richter D.J., Schaffner S.F., Gabriel S.B., Platko J.V., Patterson N.J., McDonald G.J., et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832–837.
Schlotterer C. 2002. Towards a molecular characterization of adaptation in local populations. *Curr. Opin. Genet. Dev.* **12:** 683–687.
Simonsen K.L., Churchill G.A., and Aquadro C.F. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141:** 413–429.
Sober E. 1993. *The nature of selection: Evolutionary theory in philosophical focus.* University of Chicago Press, Chicago, Illinois.
Soranzo N., Bufe B., Sabeti P.C., Wilson J.F., Weale M.E., Marguerie R., Meyerhof W., and Goldstein D.B. 2005. Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2R16. *Curr. Biol.* **15:** 1257–1265.
Stephens J.C., Schneider A.J., Tanguay D.A., Choi J., Acharya T., Stanley S.E., Jiang R., Messer C.J., Chew A., Han J., et al. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293:** 489–493.
Sunyaev S.R., Lathe W.C., III, Ramensky V.E., and Bork P. 2000. SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.* **16:** 335–337.
Suzuki Y. and Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16:** 1315–1328.
Swallow D.M. 2003. Genetics of lactase persistence and lactose intolerance. *Ann. Rev. Genet.* **37:** 197–229.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.
Tishkoff S.A., Reed F.A., Ranciaro A., Voight B.F., Babbitt C.C., Silverman J.S., Powell K., Mortensen H.M., Hirbo J.B., Osman M., et al. 2006. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39:** 31–40.
Thompson E.E., Kuttab-Boulos H., Witonsky D., Yang L., Roe B.A., and DiRienzo A. 2004. *CYP3A* variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.* **75:** 1059–1069.
Toomajian C. and Kreitman M. 2002. Sequence variation and haplotype structure at the human *HFE* locus. *Genetics* **161:** 1609–1623.
Toomajian C., Ajioka R.S., Jorde L.B., Kushner J.P., and Kreitman M. 2003. A method for detecting recent selection in the human genome from allele age estimates. *Genetics* **161:** 1609–1623.
Vallender E.J. and Lahn B.T. 2004. Positive selection on the human genome. *Hum. Mol. Genet.* **13:** 245–254.
Verrelli B.C., McDonald J.H., Argyropoulos G., Destro-Bisol G., Froment A., Drousiotou A., Lefranc G., Helal A.N., Loiselet J., and Tishkoff S.A. 2002. Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. *Am. J. Hum. Genet.* **71:** 1112–1128.
Voight B.F., Kudaravalli S., Wen X., and Pritchard J.K. 2006. A map of recent positive selection in the human genome. *PLos Biol.* **4:** 446–458.
Wall J.D. 1999. Recombination and the power of statistical tests of neutrality. *Genet. Res.* **74:** 65–79.
Wang E.T., Kodama G., Baldi P., and Moyzis R.K. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci.* **103:** 135–140.
Watterson G.A. and Guess H.A. 1977. Is the most frequent allele the oldest? *Theor. Popul. Biol.* **11:** 141–160.
Weir B.S., Cardon L.R., Anderson A.D., Nielsen D.M., and Hill W.G. 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* **15:** 1468–1476.
Wooding S.P., Watkins W.S., Bamshad M.J., Dunn D.M., Weiss R.B., and Jorde L.B. 2002. DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the *CYP1A2* gene: Implications for human population history and natural selection. *Am. J. Hum. Genet.* **71:** 528–542.
Wooding S., Kim U.K., Bamshad M.J., Larsen J., Jorde L.B., and Drayna D. 2004. Natural selection and molecular evolution in PTC, a bitter taste receptor gene. *Am. J. Hum. Genet.* **74:** 637–646.
Wong W.S., Yang Z., Goldman N., and Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168:** 1041–1051.
Yang Z. 2002. Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* **12:** 688–694.

# 25 拟 南 芥

Yan Li[1] and Justin O. Borevitz[1,2]

*[1]Department of Ecology and Evolution and [2]Committee on Genetics, University of Chicago, Chicago, Illinois 60637*



## 简介

在过去的 20 年，拟南芥一直是人们进行遗传学、发育生物学及生理学研究的模式植物。地理分布的广泛性及可利用的群体基因组工具使得拟南芥成为现代生态学和进化研究所选择的生物。全基因组序列及饱和的分子标记使得人们能清楚地区分拟南芥的不同生态型和新种。可以利用的资源和工具包括数以千计的野生新种、多套重组近交系（RILs）、显示基因（Schmid et al. 2005b）和基因组区域的（Ecker 2004；Stol et al. 2005）的基因表达图谱、一套将近完成的基因敲除突变系（Alonso et al. 2003），及一个 250K 单核苷酸多态性（SNP）的 tiling 微列阵（芯片）（我们目前的工作）。尽管拟南芥是一个近乎于自交的物种，还是有足够的杂交发生使之可以通过连锁不平衡性（LD）开展精细遗传作图。与人类相似，拟南芥的连锁不平衡性在 25～50kb 衰退（Nordborg et al. 2005）。但是，拟南芥有较高的 SNP 密度。拟南芥群体结构在局部范围内是真实的，并且在区域性和世界范围内依然存在。目前的两项利用一核心系列新种，跨越地理水平的研究，需要一个很好的统计模型来控制用于连锁作图相关的群体结

构。未来几年，这些工具的使用将会揭示拟南芥自然变种的遗传和基因组基础，并因此加深我们对遗传变异种在自然选择下的生态功能的理解。

## 拟南芥的遗传及物理作图

拟南芥第一个遗传图在1983年发布的（Koornneef et al. 1983），促进了拟南芥在20世纪80年代早期成为模式植物（Meinke et al. 1998）。之后，人们通过RIL获得到若干套遗传图谱（Lister and Dean 1993；Liu et al. 1996；Alonso-Blanco et al. 1998；Wilson et al. 2001；Louder et al. 2002，Weinig et al. 2003；Clerkx et al. 2004；Symonds et al. 2005；Werner et al. 2005）。这些图谱对绘制数量性状基因座（QTL）和精细作图是非常有用的。

拟南芥是植物界第一个完成全基因组测序的物种。有报道称，测序种哥伦比亚（Col）基因组全长为125Mb（2000年最初的拟南芥基因组），在115.4Mb的测序区域中含有25 498个基因，这个数字曾被拟南芥信息资源（TAIR）重新评估（Wortman et al. 2003），现在基因个数已被更新到31 407个（TAIR6，2005年11月11日）。另一种生态型Landsbergerecta（Ler），由Cereon Genomics以大约两倍的覆盖率（2x coverage）完成测序（http://www.arabidopsis.org/brows/Cereon/index.jsp）（Jander et al. 2002）。1267个扩增片段长度多态标记（AFLP）被定位在拟南芥基因组序列上（Peters et al. 2001）。近来，有来自96个拟南芥种的876个短片段也完成测序（晚些时候，还将有1238短片段被释放，http://walnut.usc.edu/apache2-default/2010/index-old.html）(Nordborg et al. 2005）。这种来自不同变种的测序信息加速了拟南芥遗传和基因组变异方面的研究。例如，更多的多态性标记得到了确定，这将缩短精细作图和基因克隆（Borevitz and Chory 2004）的时间，并且密集的分子定位标记使生物学家对拟南芥基因组范围内的单倍型及群体结构有更深刻的认识（Nordborg et al. 2005；Schmid et al. 2006）。

### 多态性标记与遗传图

除了传统的，如RFLP分子标记外，AFLP等显性酶切扩增多态性序列（codominant cleaved amplified polymorphic sequences，CAPS）、数以千计的SNP和插入/删除片段也都是可以利用的（Jander et al. 2002）。自从Col和Ler的基因组序列公开使用以来，Cereon Genomics预计在Col和Ler间共有56670个多态性位点（包括37 344个SNP、18 579个插入/删除片段和747个大插入/删除片段），这些数据现在可以通过网址http://www.arabidopsis.org/browse/cereon/index.jsp共享。与微卫星分子标记相比，SNP和插入/删除片段有以下几点优势（Torjek et al. 2003）：高丰度、多半共显性、通常是中间表型（Berger et al. 2001）。另8051个在12种拟南芥生态型中发现的SNP也在Max-Planck拟南芥SNP协会（MASC）的SNP数据库公布，并公开使用。近来，通过对96个种拟南芥的876个短片段的重新测序，超过17 000个SNP和插入/删除片段已经被确认（Nordborg et al. 2005）。这些多态性已经被提交到拟南芥信息资

源库（TAIR） (http://www.arabidopsis.org/news/ monthly/TAIR _ News _ Sept04.jsp) 并且可以通过网址 http://walnut.usc.edu/2010l 浏览或 http:// msqt.weigelworld.org/网站查询(N. Warthmann, unpubl.)。

Affymetrix 基因芯片微阵列为发现多态性标记和对其进行基因分型提供了高通量的研究平台。高密度寡核苷酸阵列含有数百万个 25-mer 的探针。当阵列与标定的全基因组 DNA 杂交时，每种探针都是潜在的标记。在 Col 和 Ler 间大约 4000 个单动能多态 (SFP) 位点被检测出来，其错误率是 5%左右；与 Col 生态型相比，使用仅有 100 000 个探针的阵列就在 5 个拟南芥种中检测出 4 713 个 SFP 作为种间的多态性 (BOrevitz et al. 2003)。Borevitz 和 Nordborg 实验室最近制备了一种新的阵列，整合了250 000个非单一性 SNP (R. Clark and P. Weigel in prep) 和 150 多万个 tiling 阵列探针（这种阵列能够探测新的 SNP 及插入/删除片段）。此外，来自一些植物病原体序列（包括 *Pst*、*Psm*、*Psy*、*Psx*、Agrobacteria、Xanthomonas、H. parasitica 和很多病毒）的基因及 tiling 探针也包含在这个新的整合阵列。

## 拟南芥的地理及群体结构

在早期研究中拟南芥的遗传多样性检测只能利用数量有限的分子标记，包括同种异型酶 (Abbott and Gomes 1989)、RFLP (Bergelson et al. 1998)、AFLP (Miyashita et al. 1999; Sharbel et al. 2000) 和微卫星 (Kuittimen et al. 1997)。随着拟南芥基因组内高密度分子标记的发展，两个研究组近来对全球范围内拟南芥群体的遗传变异情况进行了调查 (Nordborg et al. 2005; Schimid et al. 2006)。这两项研究根据拟南芥不同地理区域的遗传多样性差异，证明拟南芥有显著的全球群体结构 (Nordborg et al. 2005; Schimid et al. 2006)。但是，不同群体间有许多共同的多态性 (Nordborg et al. 2005)。地方的局部性群体内及群体间也有许多遗传变异 (Nordborg et al. 2005; Bakker et al. 2006)，这是对最早认为拟南芥是以无性小群体存在的论点的有力反驳。一些个别群体还含有许多种系范围内存在的变异 (Nordborg et al. 2005)。人们发现美国中西部拟南芥群体的异质性，这点可由与其他地理区域群体共有的广泛的单倍型证明 (Nordborg et al. 2005)。另也有研究表明，欧洲中部及东部的拟南芥种属于混合性群体 (Schmid et al. 2006)。

关于分子多态性模式，由于过量罕见等位基因的存在，等位基因频率的分布与标准中间模型并不相符 (Nordborg et al. 2005; Schmid et al. 2006)。分子多态性水平为一千个碱基对平均有大约 7 个核苷酸（核苷酸多样性约为 0.007）(Nordborg et al. 2005; Schmid et al. 2005a)，但是在不同的基因组区域是有波动的，这通常与部分重复区域呈正相关而与基因密度呈负相关 (Nordborg et al. 2005)。不同群体间多态性的水平也有所变化。例如，相对伊比利亚半岛及欧洲中部，中亚拟南芥群体的多态性水平是比较低的 (Schmid et al. 2006)。因此，人们认为样品采集策略和采用的分子标记对所观察到遗传变异模式有影响。

## 探索自然变异遗传基础的工具

拟南芥中仅有不到10%基因的功能被成功鉴定出来（Ostergaard and Yanofsky 2004），并且主要是通过基因敲除的实验室突变技术实现的。这一策略在今后的基因功能研究中仍然很重要，但是自然变异作为新基因发现和功能基因组学研究方法正在重新受到关注。自然变异能够克服实验室诱导突变的限制，如无明显表型变异、表型不稳定、遗传背景有局限性及表型变异范围有限等（Tonsor et al. 2005）。自然变异还有许多优势，特别是在发现新基因和新等位基因、在大遗传背景和环境中描述基因功能特性以及理解复杂性状的遗传构成等方面。

自然发生变异的基因已有一些被鉴定出来：其中抗病性相关基因有 *RTM1*、*RPS2* 和 *RPM1*（Mindrinos et al. 1994；Stahl et al. 1999；Chisholm et al. 2000）；开花相关基因包括 *FRI*（Johanson et al. 2000；Gazzani et al. 2003）、*FLC*（Gazzani et al. 2003；Michaels et al. 2003）、*EDI*（El-Assal et al. 2001）、*FLM*（Werner et al. 2005）；下胚轴长度/光照应答相关基因有 *PHYA*（Maloof et al. 2001）和 *PHYD*（Aukerman et al. 1997）；以及一些其他基因（Koornneef et al. 2004）。在大多数情况下，种间的自然变异是多基因的（Koornneef et al. 2004）。通常说来，QTL分析是进行这种自然突变的遗传基础分析的第一步。RIL已被证明是进行QTL分析的有效手段，它可以使用多个复制片段来测量表型，增强QTL的检测能力（Koornneef et al. 2004）。利用RIL已经进行了拟南芥许多性状的分析，包括对生物因子的抗性/耐受性、发育特性、生理特性、化学成分和酶的活性（Koornneef et al. 2004）。最常用的 Ler×Col RIL 群体被用来进行许多性状的表达QTL（eQTL）作图（DeCook et al. 2006）。Ler×Cvi RIL 群体已被用来进行40多个表型性状的遗传分析（Koornneef et al. 2004），并被用于2000多个代谢物的分析（Keurentjes et al. 2006）。在同一群体中同时对多个性状进行作图可以帮助比较QTL图谱的位置。近来，有9个不同的RIL群体被公开使用（http://www.arabidopsis.org/abrc/catalog/recombiant_inbred_set_1.html），但是仍有更多的群体在构建之中（http://naturalvariation.org/genetic.html，http://www.inra.fr/internet/Produits/vast/RILs.htm）。

混合分组分析法是一种基因作图或QTL作图的高效快速方法。通过与来自极端表型品系的混合DNA进行微阵列杂交，效应大的突变，如较大的缺失，可以被快速定位（Borevitz et al. 2003；Hazen et al. 2005a，b；Werner et al. 2005）。近来，利用LD作图来确定自然变异的相应基因座位的可行性已经在拟南芥进行了研究（Nordborg et al. 2002，2005；Schmid et al. 2003，2005a，2006；Olsen et al. 2004；Aranzana et al. 2005；Rosenberg and Nordborg 2006；Zhao et al. 2007）。在像拟南芥这样的自交物种中LD作图是非常有用的，因为作图群体基因分型的核心可以被许多研究人员在进行不同的遗传性状的分析时重复使用。近来的研究显示，一旦人们能控制群体结构并建立一个适合的统计模型，LD作图将成为检测自然变异遗传基础的有力工具。我们近来正在构建几套大的核心种群系列（>384），并将利用250K SNP微阵列对它们进行基

因分型以作为高分辨率 LD 作图的共享平台。

在近等基因系（NIL）或异质近交系（HIF）中可以对 QTL 进行验证和精细作图（Borevitz and Chory 2004）。多态性标记的密度及重组结果是精细作图的关键因素。在有些情况下，QTL 可以直接作图到基因上（Fridman et al. 2000；Kroymann et al. 2001）。通常，QTL 一旦被定位在一个相对小到 3cM 或更小的区域，就可以开始筛选候选基因了（Borevitz and Chory 2004）。人们可以依据多态性信息来筛选候选基因，如基因的改变／缺失或基因表达谱。候选基因一旦被确定，人们就可以利用突变体(http://signal.salk.edu)(Sessions et al. 2002;Alonso et al. 2003）和创建转基因系来进行功能研究了（Jander et al. 2002)。

# 高通量基因型分析、表型分析和相关研究方案

随着分子标记的发展及拟南芥野生变种的收集，基因分型和表型分析都需要高通量的方法。消费成本是一个小实验室研究自然变异要考虑的另一个因素。在我们实验室中，一直在使用高产率的方法提取 DNA，进行基因分型和表型分析，但是到目前为止表型确定仍然是一个限制步骤。以下列出的是研究方案和资源。

## DNA 提取

PUREGENE（Gentra System）的基因组 DNA 纯化试剂盒是高质量、高产率 DNA 提取的一个选择，我们用此方案提取 DNA 进行 SNP 基因分型得到了满意的结果。我们将提取方案修改成适合 96 孔板的模式（第 7 章）。

## SNP 基因分型和微阵列基因分型

我们使用了 149 个 SNP 标记，以中等等位基因的频率来为带有不同遗传背景的 $F_2$ 杂交群体作图(www.naturalvariation.org.)。SNP 基因分型是通过 Sequenom 在 iPLEX 库中（每个库大约 40 个 SNP）分析完成的。利用这些 SNP，我们已经检测了 5000 个以上变种（包括来自 Columbus Ohio 的拟南芥生物资源中心的 800 多个变种），确定的基因型在我们的网站可以共享 www.naturalvariation.org。微阵列基因分型方法在之前已叙述（Borevitz et al. 2003；Werner et al. 2005）。

## 表型分析

我们使用条形码技术来记录植物表型（图 25-1）。每个变种的图片是在植株开花时照的（第一批花芽出现），这些照片可以从网站 www.naturalvariation.org 获得。条形码阅读软件（Softek Softeat,www.barcode.com）用来重新命名带有条形码名字（修改版可以从网站 www.naturalvariation.org 获得）的照片。2007 年，我们和其他工作者将会开始使用定时摄影技术来记录发育程序变异的图谱。我们可以用形态测量分析来量化表型变化范围。

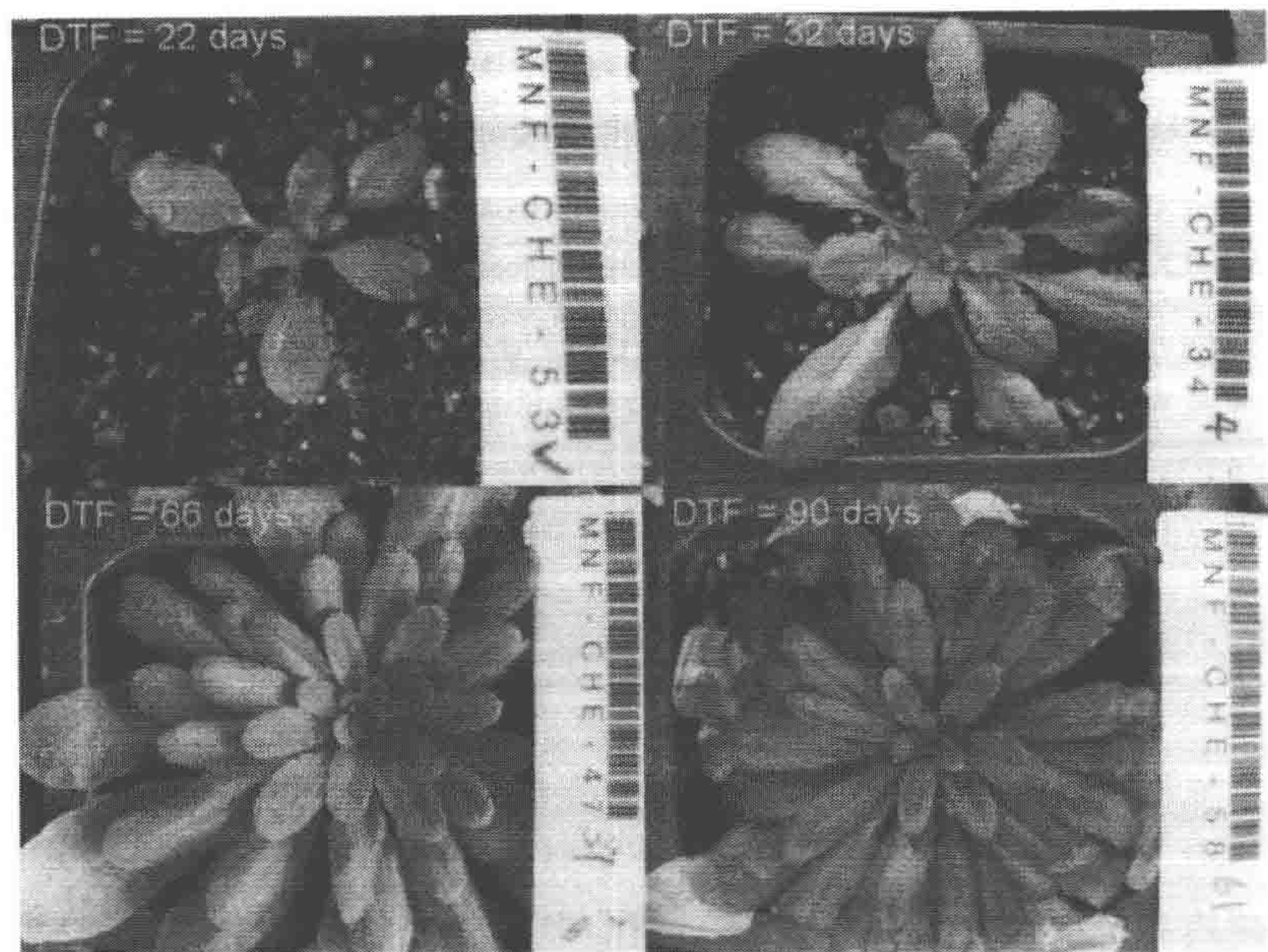



图 25-1　某地区拟南芥开花时间的自然变异（MNF-CHE 在美国中西部种群）。DTF（Days to flower），开花天数是指从种子播种到土壤中开始到第一批花芽的出现。标记是使用的条形码技术

# 资源

表 25-1 列出了有关拟南芥野生种系收集、序列、多态性及突变体的数据库和网址。表 25-2 列出了公开可以利用的 RIL。

**表 25-1　研究拟南芥自然变异的资源**

| 网站数据库 | 主要作用 |
|---|---|
| http://www.arabidopsis.org/ | 序列、多变性、种子 |
| http://msqt.weigelworld.org/ | Nordborg 96 个变种的 SNP 的多态性 |
| http://walnut.usc.edu/2010 | |
| http://www.arabidopsis.org/ | |
| http://www2.mipz-koeln.mpg.de/masc/ | |
| http://www.arabidopsis.org/browse/Cereon/index.jep | 12 种生态型中的 SNP 多态性 |
| http://www.naturalvariation.org | 在 Col/Ler 中的多态性 |
| http://signal.salk.edu | 野生型的收集和它们的图片、基因型、SNP、微阵列基因型分析方法等 |
| | 突变体 |

**表 25-2 通过 TAIR 公共资源可获得的 RILs**

(http://www.arabidopsis.org/abrc/catalog/recombinant_inbred_set_1.html)

| 亲本型（父本×母本） | 捐赠者 | 特征/用途[a] |
|---|---|---|
| Col-3×Nd-1 and<br>Col-5 (gl1) ×Nd-1 | Eric Holub<br>Jim Beynon<br>Ian Crute | 抗病能力 |
| Van-0×Col-0 (WT-2) | Justin Borevitz<br>Evadne Smith<br>Julin Maloof<br>Detlef Weigel<br>Joanne Chory | 花期 |
| Ler-0/No-0[b] | Jorge Casal<br>Javier Botto<br>Alan Lloyd | 胚轴的发育 |
| Bay-0×Sha | Olivier Loudet<br>Sylvain Chaillou | 氮的使用效率、光反应、根的生长、碳水化合物的含量、细胞壁的组成 |
| Ler-0×Col-4 | Clare Lister<br>Caroline Dean | 改进基因和物理图谱 |
| Ler-0×Sha | Maarten Koornneef | 花期 |
| Cvi-1/Ler-2[b] | Maarten Koornneef | 抗虫能力、耐寒能力、花期、植株和种子的大小、种子的休眠等 |
| Ler×Ws-1 | Pablo Scolnik | 改进基因图谱 |
| Col (gl1) ×Kas-1 | Shauna Somerville | 抗白粉病能力、花期 |

a. 信息来自 TAIR 网站；b. 杂种系谱不清楚。

# 结论

拟南芥作为模式生物体系已经有 20 多年了，所获得的大量多态性标记、高通量研究方法及野生种系的收集都将使其能继续在遗传学、发育学、生理学、生态学和进化生物学方面的研究中作出贡献。关联作图将会在拟南芥许多性状候选基因的确定，尤其是生态相关性状和那些产量/疾病抗性相关性状得到广泛的应用。在拟南芥中建立的研究方法也可以应用到其他植物体系，尤其是那些自交（自花授粉）的作物。

# 致谢

我们十分感谢 *Bergelson J* 和 *Nordborg M* 博士的合作，*Scholl R* 和 *RIvero L* 博士提供 850 分库存拟南芥核心品系，及 Holub E 和 Byers D 博士在关联作图项目中提供的拟南芥变种。我们的这部分工作得到 NIH 的资助（Grant R01 GM073822）。

# 参考文献


Abbott R.J. and Gomes M.F. 1989. Population genetic-structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. *Heredity* **62:** 411–418.

Alonso J.M., Stepanova A.N., Leisse T.J., Kim C.J., Chen H.M., Shinn P., Stevenson D.K., Zimmerman J., Barajas P., Cheuk R., et al. 2003. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301:** 653–657.

Alonso-Blanco C., Peeters A.J.M., Koornneef M., Lister C., Dean C., van den Bosch N., Pot J., and Kuiper M.T.R. 1998. Development of an AFLP based linkage map of *Ler*, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a *Ler*/Cvi recombinant inbred line population. *Plant J.* **14:** 259–271.

Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Aranzana M.J., Kim S., Zhao K.Y., Bakker E., Horton M., Jakob K., Lister C., Molitor J., Shindo C., Tang C.L., et al. 2005. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *Plos Genet.* **1:** 531–539.

Aukerman M.J., Hirschfeld M., Wester L., Weaver M., Clack T., Amasino R.M., and Sharrock R.A. 1997. A deletion in the *PHYD* gene of the *Arabidopsis* Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. *Plant Cell* **9:** 1317–1326.

Bakker E.G., Stahl E.A., Toomajian C., Nordborg M., Kreitman M., and Bergelson J. 2006. Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Mol. Ecol.* **15:** 1405–1418.

Bergelson J., Stahl E., Dudek S., and Kreitman M. 1998. Genetic variation within and among populations of *Arabidopsis thaliana*. *Genetics* **148:** 1311–1323.

Berger J., Suzuki T., Senti K.A., Stubbs J., Schaffner G., and Dickson B.J. 2001. Genetic mapping with SNP markers in *Drosophila*. *Nat. Genet.* **29:** 475–481.

Borevitz J.O. and Chory J. 2004. Genomics tools for QTL analysis and gene discovery. *Curr. Opin. Plant Biol.* **7:** 132–136.

Borevitz J.O., Liang D., Plouffe D., Chang H.S., Zhu T., Weigel D., Berry C.C., Winzeler, E., and Chory, J. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13:** 513–523.

Chisholm S.T., Mahajan, S.K., Whitham, S.A., Yamamoto, M.L., and Carrington, J.C. 2000. Cloning of the *Arabidopsis RTM1* gene, which controls restriction of long-distance movement of tobacco etch virus. *Proc. Natl. Acad. Sci.* **97:** 489–494.

Clerkx E.J.M., El-Lithy M.E., Vierling E., Ruys G.J., Blankestijin-De Vries H., Groot S.P.C., Vreugdenhil D., and Koornneef M. 2004. Analysis of natural allelic variation of *Arabidopsis* seed germination and seed longevity traits between the accessions Landsberg *erecta* and Shakdara, using a new recombinant inbred line population. *Plant Physiol* **135:** 432–443.

DeCook R., Lall S., Nettleton D., and Howell S.H. 2006. Genetic regulation of gene expression during shoot development in *Arabidopsis*. *Genetics* **172:** 1155–1164.

Ecker J.R. 2004. Genome-wide discovery of transcription units and functional elements in arabidopsis. *Mol. Biol. Cell* **15:** 122A.

El-Assal S.E.D., Alonso-Blanco C., Peeters A.J.M., Raz V., and Koornneef M. 2001. A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2*. *Nat. Genet.* **29:** 435–440.

Fridman E., Pleban T., and Zamir D. 2000. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proc. Natl. Acad. Sci.* **97:** 4718–4723.

Gazzani S., Gendall A.R., Lister C., and Dean C. 2003. Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. *Plant Physiol.* **132:** 1107–1114.

Hazen S.P., Schultz T.F., Pruneda-Paz J.L., Borevitz J.O., Ecker J.R., and Kay S.A. 2005a. *LUX ARRHYTHMO* encodes a Myb domain protein essential for circadian rhythms. *Proc. Natl. Acad. Sci.* **102:** 10387–10392.

Hazen S.P., Borevitz J.O., Harmon F.G., Pruneda-Paz J.L., Schultz T.F., Yanovsky M.J., Liljegren S.J., Ecker J.R., and Kay S.A. 2005b. Rapid array mapping of circadian clock and developmental mutations in *Arabidopsis*. *Plant Physiol.* **138:** 990–997.

Jander G., Norris S.R., Rounsley S.D., Bush D.F., Levin I.M., and Last R.L. 2002. Arabidopsis map-based cloning in the post-genome era. *Plant Physiol.* **129:** 440–450.

Johanson U., West, J., Lister, C., Michaels, S., Amasino, R., and Dean, C. 2000. Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* **290:** 344–347.

Keurentjes J.J.B., Fu J.Y., de Vos C.H.R., Lommen A., Hall R.D., Bino R.J., van der Plas L.H.W., Jansen R.C., Vreugdenhil D., and Koornneef M. 2006. The genetics of plant metabolism. *Nat. Genet.* **38:** 842–849.

Koornneef M., Alonso-Blanco C., and Vreugdenhil D. 2004. Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.* **55:** 141–172.

Koornneef M., Vaneden J., Hanhart C.J., Stam P., Braaksma F.J., and Feenstra W.J. 1983. Linkage map of *Arabidopsis thaliana*. *J. Hered.* **74:** 265–272.

Kroymann J., Textor S., Tokuhisa J.G., Falk K.L., Bartram S., Gershenzon J., and Mitchell-Olds T. 2001. A gene controlling variation in arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. *Plant Physiol.* **127:** 1077–1088.

Kuittinen H., Mattila A., and Savolainen O. 1997. Genetic variation at marker loci and in quantitative traits in natural populations of *Arabidopsis thaliana*. *Heredity* **79:** 144–152.

Lister C. and Dean C. 1993. Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.* **4:** 745–750.

Liu Y.G., Mitsukawa N., Lister C., Dean C., and Whittier R.F. 1996. Isolation and mapping of a new set of 129 RFLP markers in *Arabidopsis thaliana* using recombinant inbred lines. *Plant J.* **10:** 733–736.

Loudet O., Chaillou S., Camilleri C., Bouchez D., and Daniel-Vedele F. 2002. Bay-0 x Shahdara recombinant inbred line population: A powerful tool for the genetic dissection of complex traits in *Arabidopsis*. *Theor. Appl. Genet.* **104:** 1173–1184.

Maloof J.N., Borevitz J.O., Dabi T., Lutes J., Nehring R.B., Redfern J.L., Trainer G.T., Wilson J.M., Asami T., Berry C.C., Weigel D., and Chory J. 2001. Natural variation in light sensitivity of *Arabidopsis*. *Nat. Genet.* **29:** 441–446.

Meinke D.W., Cherry J.M., Dean C., Rounsley S.D., and Koornneef M. 1998. *Arabidopsis thaliana*: A model plant for genome analysis. *Science* **282:** 662, 679–682.

Michaels S.D., He Y.H., Scortecci K.C., and Amasino R.M. 2003. Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis*. *Proc. Natl. Acad. Sci.* **100:** 10102–10107.

Mindrinos M., Katagiri F., Yu G.L., and Ausubel F.M. 1994. The *A. thaliana* disease resistance gene *RPS2* encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell* **78:** 1089–1099.

Miyashita N.T., Kawabe A., and Innan H. 1999. DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. *Genetics* **152:** 1723–1731.

Nordborg M., Borevitz J.O., Bergelson J., Berry C.C., Chory J., Hagenblad J., Kreitman M., Maloof J.N., Noyes T., Oefner P.J., Stahl E.A., and Weigel D. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **30:** 190–193.

Nordborg M., Hu T.T., Ishino Y., Jhaveri J., Toomajian C., Zheng H.G., Bakker E., Calabrese P., Gladstone J., Goyal R., et al. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *Plos Biol.* **3:** 1289–1299.

Olsen K.M., Halldorsdottir S.S., Stinchcombe J.R., Weinig C., Schmitt J., and Purugganan M.D. 2004. Linkage disequilibrium mapping of *Arabidopsis CRY2* flowering time alleles. *Genetics* **167:** 1361–1369.

Ostergaard L. and Yanofsky M.F. 2004. Establishing gene function by mutagenesis in *Arabidopsis thaliana*. *Plant J.* **39:** 682–696.

Peters J.L., Constandt H., Neyt P., Cnops G., Zethof J., Zabeau M., and Gerats T. 2001. A physical amplified fragment-length polymorphism map of *Arabidopsis*. *Plant Physiol.* **127:** 1579–1589.

Rosenberg N.A. and Nordborg M. 2006. A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. *Genetics* **173:** 1665–1678.

Schmid K.J., Ramos-Onsins S., Ringys-Beckstein H., Weisshaar B., and Mitchell-Olds T. 2005a. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169:** 1601–1615.

Schmid K.J., Torjek O., Meyer R., Schmuths H., Hoffmann M.H., and Altmann T. 2006. Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *Theor. Appl. Genet.* **112:** 1104–1114.

Schmid K.J., Sorensen T.R., Stracke R., Torjek O., Altmann T., Mitchell-Olds T., and Weisshaar B. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* **13:** 1250–1257.

Schmid M., Davison T.S., Henz S.R., Pape U.J., Demar M., Vingron M., Scholkopf B., Weigel D., and Lohmann J.U. 2005b. A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37:** 501–506.

Sessions A., Burke E., Presting G., Aux G., McElver J., Patton D., Dietrich B., Ho P., Bacwaden J., and Ko C. 2002. A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell* **14:** 2985–2994.

Sharbel T.F., Haubold B., and Mitchell-Olds T. 2000. Genetic isolation by distance in *Arabidopsis thaliana*: Biogeography and postglacial colonization of Europe. *Mol. Ecol.* **9:** 2109–2118.

Stahl E.A., Dwyer G., Mauricio R., Kreitman M., and Bergelson J. 1999. Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*. *Nature* **400:** 667–671.

Stolc V., Samanta M.P., Tongprasit W., Sethi H., Liang S.D., Nelson D.C., Hegeman A., Nelson C., Rancour D., Bednarek S., et al. 2005. Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc. Natl. Acad. Sci.* **102:** 4453–4458.

Symonds V.V., Godoy A.V., Alconada T., Botto J.F., Juenger T.E., Casal J.J., and Lloyd A.M. 2005. Mapping quantitative trait loci in multiple populations of *Arabidopsis thaliana* identifies natural allelic variation for trichome density. *Genetics* **169:** 1649–1658.

Tonsor S.J., Alonso-Blanco C., and Koornneef M. 2005. Gene function beyond the single trait: Natural variation, gene effects, and evolutionary ecology in *Arabidopsis thaliana*. *Plant Cell Environ.* **28:** 2–20.

Torjek O., Berger D., Meyer R.C., Mussig C., Schmid K.J., Sorensen T.R., Weisshaar B., Mitchell-Olds T., and Altmann T. 2003. Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*. *Plant J.* **36:** 122–140.

Weinig C., Stinchcombe J.R., and Schmitt J. 2003. QTL architecture of resistance and tolerance traits in *Arabidopsis thaliana* in natural environments. *Mol. Ecol.* **12:** 1153–1163.

Werner J.D., Borevitz J.O., Warthmann N., Trainer G.T., Ecker J.R., Chory J., and Weigel D. 2005. Quantitative trait locus mapping and DNA array hybridization identify an *FLM* deletion as a cause for natural flowering-time variation. *Proc. Natl. Acad. Sci.* **102:** 2460–2465.

Wilson I.W., Schiff C.L., Hughes D.E., and Somerville S.C. 2001. Quantitative trait loci analysis of powdery mildew disease resistance in the *Arabidopsis thaliana* accession Kashmir-1. *Genetics* **158:** 1301–1309.

Wortman J.R., Haas B.J., Hannick L.I., Smith R.K., Maiti R., Ronning C.M., Chan A.P., Yu C.H., Ayele M., Whitelaw C.A., White O.R., and Town C.D. 2003. Annotation of the Arabidopsis genome. *Plant Physiol.* **132:** 461–468.

Zhao K., Aranzana M.J., Kim S., Lister C., Shindo C., Tang C., Toomajian C., Zheng H., Dean C., Marjoram P., and Nordborg M. 2007. An *Arabidopsis* example of association mapping in structured samples. *Plos Genet.* **3:** e4.

# 互联网资源

**http://www.arabidopsis.org/abrc/catalog/recombinant_inbred_set_1.html** Sets of Recombinant Inbred Lines, Arabidopsis Biological Resource Center (ABRC)

**http://www.arabidopsis.org/news/monthly/TAIR_News_Sept04.jsp** The Arabidopsis Information Resource (TAIR) news release

**http://www.inra.fr/internet/Produits/vast/RILs.htm** V.A.S.T. (Variation and Abiotic Stress Tolerance) Lab tries to list available and upcoming Arabidopsis RILs (or NILs), which are useful for quantitative genetics

**http://walnut.usc.edu/apache2-default/2010/index-old.html** A genomic survey of polymorphism and linkage disequilibrium, funded by the NSF 2010 Project. J. Bergelson, M. Kreitman, Department of Ecology & Evolution, University of Chicago; M. Nordborg, Program in Molecular & Computational Biology, University of Southern California

# 26 玉 米

W. Brad Barbazuk,[1] An-Ping Hsia,[2] Hsin D. Chen,[2] Yan Fu,[1]
Kazuhiro Ohtsu,[2] and Patrick S. Schnable[2,3,4]

*[1]Donald Danforth Plant Science Center, St. Louis, Missouri 63132; [2]Department of Agronomy, [3]Department of Genetics, Development, and Cell Biology, [4]Center for Plant Genomics, Iowa State University, Ames, Iowa 50011*



## 简介

玉米既是全球性的重要作物，又是基因组结构和功能研究的经典体系。玉米在营养学、农业和工业上的用途促使各项针对玉米基因改造的研究项目的持续开展。而近年来对安全可再生生物能源的大力搜寻是推动对这个重要作物进行彻底的生物学和遗传学研究的又一强劲动力。已有的古生态学和分子生物学数据表明，玉米是 9000～6000 年前由 Zea mays ssp. Parviglumis 驯化而来的（Piperno and Flannery 2001；Matsuoka et al. 2002）。玉米及其近亲蜀黍在形态学水平上有多方面的不同，控制这些性状的基因都是驯化玉米的土著美国人所选择的目标，植物育种专家进一步巩固了满足人类需求的特征。这样，玉米也成为研究驯化与人工选择的极好模式植物。

玉米基因组大，据估算有多达 50 000 个基因分布在 10 条染色体上，使得单倍体基因组大小达到近 25 亿个核苷酸对。玉米基因似乎是小基因组簇形式贯穿分布在基因组中，被大段的重复 DNA 所分开（Martienssen et al. 2004；Barbazuk et al. 2005；Rabinowicz and Bennetzen 2006）。玉米基因组整合数据库和基因组工具，如遗传图谱（Cone et al. 2002；Lee et al. 2002；Sharopova et al. 2002；Fu et al. 2006）和物理图谱（Coe et al. 2002；Cone et al. 2002）、包括 100 万以上表达序列标签（EST）

片段序列和超过 100 万的基因组勘测序列（GSS）的集合数据库。一个由公共基金支持对玉米自交系 B73 的全基因组测序项目也在进行中（Rabinowicz and Bennetzen 2006）。

玉米的遗传多样性十分丰富，在自交系和大田品种之间的 SNP 和插入/缺失多态性（IDP）频率是在编码区平均每 124 个碱基有一个（Ching et al. 2002），在全部相关区域则是每 28 个碱基有一个（Tanaillon et al. 2001）。异常丰富的遗传多样性成为表型多样性的基础，正是这种表型多样性使得玉米可以广泛适应不同环境并表现出很强的杂种优势。杂种优势，或者说"杂合活力"是由偶然出现的基因和等位基因之间的互作引起的优良特性的表现。关于杂种优势机理和遗传多样性作用的深入理解和认识对未来植物育种工作的成败起着至关重要的作用。这种多样性很容易由遗传学多态性加以确认，包括 IDP 和 SNP，这些可以转化为遗传标记的多态性并被高通量低成本地检出（Gut 2001；Kwok 2001）。而正是由于这些多态性的大量存在，使基于 IDP 和 SNP 标记构建高精度遗传图谱成为可能（Rafalski 2002）。这些图谱可以用来进行基因组结构组成和功能的研究，还可在标记辅助选择项目中加以应用和促进有关进化和减数分裂重组基本问题的研究。IDP 和 SNP 还能进一步应用在利用基因组范围连锁不平衡和基因组关联研究中有特殊功能或性状的相关基因的研究。此外，转录关联的 SNP 可以用于在玉米顺式调控变异的检验中进行等位基因特异性分析（Cowles et al. 2002；Bray et al. 2003；Guo et al. 2004；Pastinen et al. 2004；Stupar and Springer 2006）。

本章，我们讨论了玉米遗传图谱研究的历史和构成及其目前遗传图谱的研究状况。而且，我们提出了一项快速构建玉米 IDP 标记用来增加遗传图谱分子标记密度的策略。高频率的 IDP 标记可以利用高通量技术进行检测。重要的是，以 IDP 标记的多态性在作图谱前无需鉴定。在本章中我们也将简要讨论一些在本书其他章节中出现的一些方法：植物冻干组织的 DNA 提取（第 7 章）；从植物组织中获取不连续细胞群，并从中分离和扩增 RNA（第 8 章）；利用温度梯度毛细管电泳检测多态性（TGCE）和 SNP 数据的挖掘（第 10 章）。我们还要介绍由生命科学 454 发明的新的海量平行 DNA 序列检测平台（Margulies et al. 2005），并解释了这个平台如何用于快速检测玉米的转录关联的 SNP。

## 玉米分子遗传图谱

遗传学图使得在未知遗传变异的表型效应情况下的数量性状基因座作图或基因作图提供了可能。首次发表的玉米遗传图（Emerson et al. 1935）仅仅包括 62 个位点，每个位点都是通过一个形态学突变体确定的。从那时起，多重遗传图谱就诞生了。首个玉米分子图是利用 RFLP 分子标记和一个 $F_2$ 群体构建的（Helentjaris et al. 1986），而 Burr 等（1988）绘制了一套重组自交系（RIL）的 RFLP 图谱。Davis 等（1999）利用核心本标记（Core bin markers）、EST、RFLP 和序列标签位点（STS）标记构建了玉米的遗传图谱，他们所用的作图群体来自 Tx303 与 CO159 杂交产生的 54 个永久 $F_2$ 群体。这个图谱被称为 UMC98 图谱，已经成为植物科学界的众多原创工作的一种非常有

用的核心资源。然而，由于它分辨率的限制和对 RFLP 标记印迹杂交的依赖性，阻碍了它在高通量遗传作图中的应用。

Lee 等（2002）通过自交系 B73 和 Mo17 的一个单交种的 $F_2$ 种群的随机互交，构建了一个 B73×Mo17（IBM）的互交群体，这个 $F_2$ 种群在相互交配重组近亲繁殖系产生之前已繁殖了好几个世代。由于互交过程的后代可以提供更多的重组机会，因此这样的作图群体可以使鉴定率大大增加。利用这种 IBM 群体和可用的玉米 DNA 序列资源，超过 2300 个以 PCR 为基础的标记，也就是说 1000 多个微卫星标记（Sharopova et al. 2002）和 1300 多个 IDP 标记（Fu et al. 2006）被发现，完成作图。通过互联网（http://magi. plantgenomics. iastate. edu）可以访问 ISU（爱荷华州立大学）的-IDP 图谱、IDP 引物序列、设计 IDP 引物的原初序列、优化的特定标记的 PCR 条件以及所有 IDP 标记的多态性数据。通用的标记框架是从其他图谱（SSR T218×GT119，SSR Tx303×Co159，UMC98，BNI96 框，BNL2002）的数据（Cone et al. 2002），外加 5400 个非框架标记（http：//www. maizemap. org）整合而来的。IBM 系种子来自于玉米遗传学合作种质中心（Urbana，Illinois），IBM 分组的 94 个系的 DNA 以 96 孔板的形式可以公开使用（http：//www. maizemap. org/dna _ kits. htm）。近来广泛应用的玉米分子遗传图谱可以通过互联网接口进入玉米遗传和基因组学数据库（MaizeGDB；http：//www. maizezegdb. org）。表 26-1 总结了本章及全书有关玉米研究涉及的互联网地址和软件包。

**表 26-1 软件和网络资源**

| 软件包 | |
|---|---|
| POLYBAYES | http://bioinformatics. bc. edu/marthlab/polybayes. html |
| GeneSeqer | http://deepc2. psi. iastate. edu/cgi-bin/gs. cgi |
| GMAP | http://www. gene. com/share/gmap |
| 网 络 资 源 | |
| http://www. maizemap. org | 玉米作图项目网站 |
| http://www. maizezegdb. org | 玉米基因组数据库(玉米遗传学和基因组学数据库) |
| http://www. tigr. org/tdb/e2k1/osa1 | 基因组研究所(TIGR)水稻基因组注解数据库 |
| http://compbio. dfci. harvard. edu/tgi | 基因索引入口(部分植物动物微生物 EST 的集合) |
| http://www. plantgdb. org | 植物基因组数据库(植物比较基因组学资源) |
| http://www. 454. com | 454 生命科学网站(包括技术平台和测序服务中心链接) |
| http://magi. plantgeomics. iastate. edu | 玉米基因组序列整合和其他的玉米序列资源 |
| http://www. sequenom. com | 连接技术平台的 Sequenom 主页 |
| http://www. agcol. arizona. edu | 亚利桑那州基因组学计算实验室(连接玉米序列数据和 SyMAP,以及同线作图和分析方案) |
| http://www. panzea. org | Panzea 网站:玉米的分子和功能多样性 |

一个基因的作图，要求该基因的等位基因多态性确定并能够在作图群体中被检测出来。科学家们能够利用各种技术检测多种类型的 DNA 多态性（Kristensen et al. 2001）。检测技术的选择依赖于是否提前知晓 DNA 序列与多态性之间的联系。已经通过序列分析预先检测鉴定的 SNP 或者小 IDP 可以利用相关平台进行作图，这些平台包括焦磷酸测序技术、PCR、微阵列和质谱分析技术（如 Sequenom 公司的服务）。

作为替代技术，异源双链分析可以检测到未知序列的SNPs或者小IDP，这些多态位点在定位群体的亲本中存在但是多态性的序列基础并不知道。高频的IDP多态性可以通过PCR引物扩增3′非翻译序列（UTR）或跨越内含子的PCR来检出。引物设计的候选区域可以通过下列策略鉴定，这些策略可以运用于任何植物或动物体系。

## 获取IDP的引物设计策略

（1）如果EST或者基因组序列数据是可利用的，可以通过序列比对来确定外显子。在http://magi.plantgenomics.iastate.edu/等网站可以获得部分整合的玉米基因组序列。GeneSeqer（Brendel et al. 2004）和GMAP（Wu and Watanabe 2005）都是进行EST与基因组序列比对的有效拼接工具，并且是免费的。

（2）当只有EST序列时，可以通过已知的ESTs与其他相关物种的基因组序列的比对来预测基因结构（Wei et al. 2005），如水稻（Oryza sativa spp. Nipponbarre）基因组整合的数据可以在基因组研究所（TIGR；http://www.tigr.org/tdb/e2k1/osa1）使用。GeneSeqer(Brendel et al. 2004）和GMAP（Wu and Watanabe 2005）都可以进行种间交叉和外显子拼接的比对分析的操作。

（3）当只有基因组序列数据时，可以通过已知的基因组序列在与其他相关物种（如水稻、高粱、小麦、大麦）的EST序列的比对来预测基因结构。TIGR基因索引（Quackenbush et al. 2001；http://compbio.dfci.harvard.edu/tgi/）植物基因组数据库（Dong et al. 2005；http://www.plantgdb.org）为我们提供了所收集的EST和许多植物物种的EST数据。

## 多态性分析

冻干植物组织DNA的提取可使用一种溴化十六烷基三甲铵（Rogers and Blendich 1985）的DNA提取方法的修正版本（Dietrich et al. 2002）在96孔板中进行。TGCE法可以用来分析DNA多态性，因为它可以提供一个高通量的异源双链多元分析的平台。TGCE每天可以运行12块96孔板，如果两个检测都是多元化的则至少能获得2304个数据点（Hsia et al. 2005）。这种敏感的检测分析能探查出一个800 bp的扩增子中仅有的一个SNP，或在相当于500 bp的扩增子只有1 bp IDP的多态性（Hsia et al. 2005）。该方法能准确可靠地检出SNP简单重复序列，（SSR）和IDP（Hsia et al. 2005）。所获数据可以利用Revelation软件（2.4版本）进行记录和整理分析，还有一个叫做遗传重组分析作图助手（GRAMA）的新软件包，可以促进TGCE数据的分析（Maher et al. 2006）。GRAMA能将作图数据直接显示出来可减少手工操作过程，也能通过运用替代算法从软件输出有效的数据，软件可以在http://www.complex.iastate.edu/download/GRAMA/index.html网站下载使用。Revelation软件包产生的电泳图谱在多态性缺失的位点产生一个峰，否则的话产生多个峰。例如，在图26-1，MAGI _ 65972的来自12个玉米IBM（B73XMo17 ）IRIL PCR扩增子(B73和Mo17之间有多态性)，与来自自交系B73扩增子混合（Lee et al. 2002）。图26-1左上角的插入物显示预期的同态性单一峰图模式（只有B73等位基因）和二态性

（含有 B73 和 Mo17 两个等位基因）的 PCR 混合物。因此，就可以根据电泳图谱的峰图模式确定每一个重组自交系（RIL）的基因型（图 26-1 上样孔位置表示的基因型标记）。

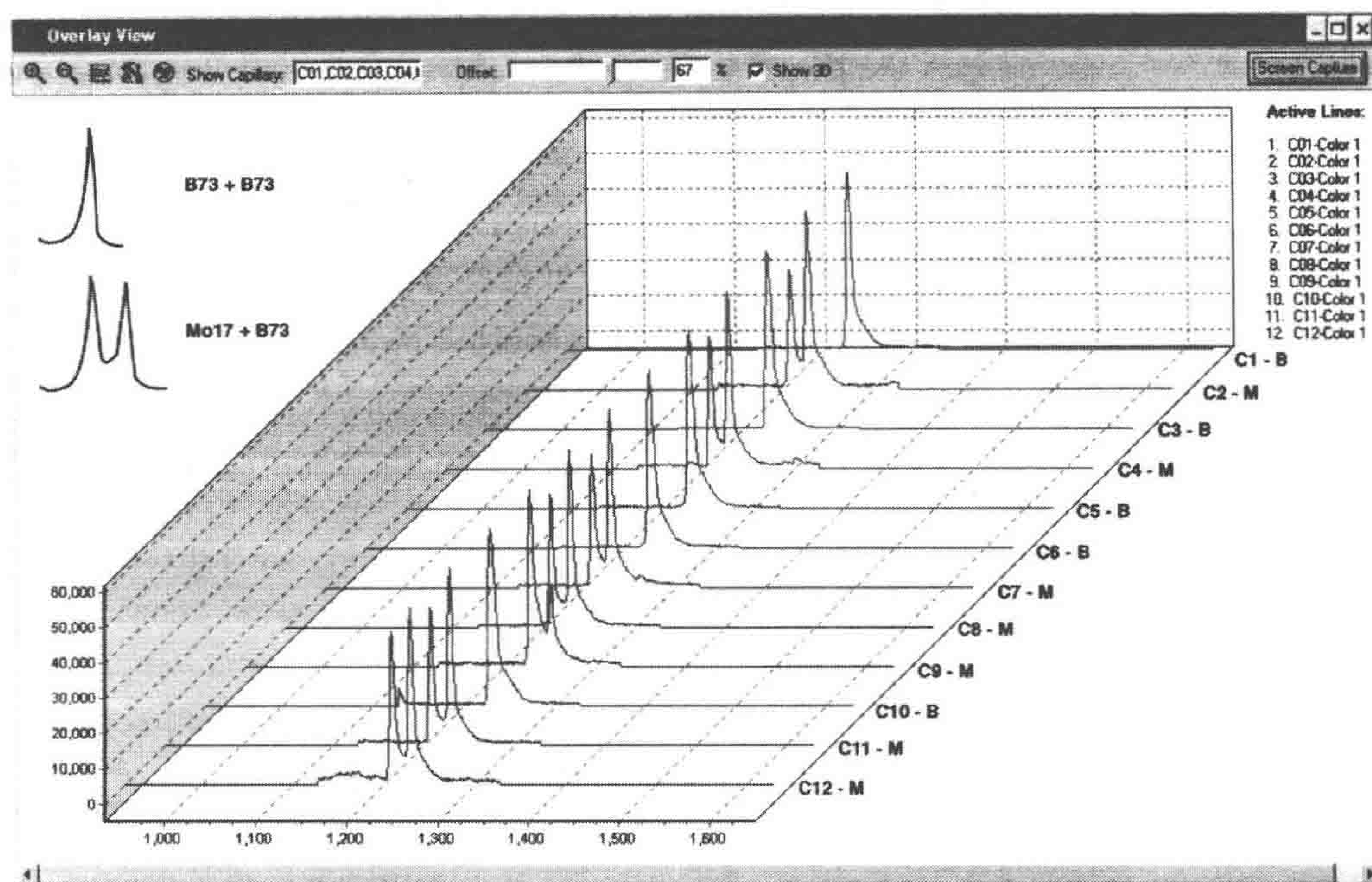



图 26-1　利用引物 31bMAGI _ 65972L5 和 31bMAGI _ 65972R4，通过 TGCE 技术对 MAGI _ 65972 进行的基因分型。左上角的标志显示预期的来自 B73 和 Mo17 等位基因的峰图。上下图分别是来自自交系 B73（标记为“B”）和 Mo17（标记为“M”）的扩增产物分别与 B73 扩增产物的混合后的电泳结果。注意代表异型双链的是来自 Mo17＋B73 的下图（多峰），而不是仅来自 B73 扩增产物的上图。C1-C12 标志来自 RIs M0034，M0039，M0043，M0045，M0048，M0052，M0055，M0058，M0061，M0063，M0067 和 M0075 的扩增子与自交系 B73 扩增产物的混合。电泳图谱显示出杂合双链的证据（多峰图）来自带有 Mo17 等位基因的 RIs，而那些单峰来自于 B73。每一个 RIs 的基因型（B 和 M）标记（位于每一上样孔的右侧）是由 Revelation 软件自动产生的。引物序列：31bMAGI _ 65972L5：ACGAGAGCTGCAATCGAATC；31bMAGI _ 65972R4：TTAG-GTGGTGGTTCGTCTCC

## 用于检测 SNP 的玉米转录组高通量 454 测序技术

基于 SNP 的标记可以通过来自两个以上玉米株系的基因组序列（Yamasaki et al. 2005）或 EST 序列（Batly et al. 2003）的比较被探测出来。高通量测序技术的最新进展提供了一个快速、廉价的 DNA 序列检测分析平台，可用于从多个玉米品系中 SNP 的检出。生命科学 454 开发出了一个可分级的高度平行的 DNA 测序系统，该系统测序速度比标准测序方法快 100 倍，目前可有能力达到每运行 4h 完成 200 000 个 DNA 片段的测序（Margulies et al. 2005）。尽管 454 测序相对较短，但这并未给基因组 DNA 测序和 454EST 与基因组模板比对比较带来问题，454 技术还被应用在转录组的

序列分析（Baibridge et al. 2006；Emrich et al. 2006；Gowda et al. 2006）。这一应用非常适合，因为转录组序列明显小于它所来源的基因组序列，含有少的重复 DNA，并且 454 也不需要构建基于克隆的文库。激光捕获显微切割技术（LCM，Schnable et al. 2004）应用到在转录子富集细胞的转录子的分离提取，这样的转录子用于 454 测序能进一步缩减目标转录组的大小，使之更有利于捕获稀有转录本序列。必要时还可以通过扩增来增加所收集的 RNA 的量。LCM-454 转录组测序在玉米中已经得到了证明（Emrich et al. 2006），在自交系 B73 和 Mo17 玉米重要的茎尖分生组织细胞（SAM）提取的 cDNA（Baurle and Laux 2003；Guyomarch et al. 2005）已成功地用于 SNP 的检测分析（W. B. Barbazuk et al.，unpubl.）。

## SNP 挖掘

假定的 SNP 可通过序列比对的错配来证实。关于 SNP 鉴定，有一些计算工具可供选择使用（Nickerson et al. 1997；Marth et al. 1999；Manaster et al. 2005；Wang and Huang 2005；Weckx et al. 2005；Zhang et al. 2005）。POLYBAYES 已经被用在包括玉米（Useche et al. 2001）的几个生物系统检测 SNP，并且被推荐用于 454 序列的 SNP 检测分析。POLYBAYES 运用贝叶斯统计模型，这种模型考虑覆盖度、序列质量及预期的多态性比例来决定多态性位点的可能性。这种可能性是指在多重序列比对（MSA）中的 SNP，而不是来自其他序列错误或者平行位片段序列比对（不是等位基因）引起的差异（Marth et al. 1999）。POLYBAYES 的分析结果可以输出易操作的 text 文件，或能够在 CONSED 中可以显示图形和进行阅读的 .ace 文件（Gordon et al. 1998）。另外，POLYBAYES 利用不管是基因组还是 EST 序列数据作模板，在多态性扫描分析之前，可以依赖这种模板将所有其他序列通过 CROSS _ MATCH（P. Green，unpubl.）进行多重序列比对来进行评定。以模板为基础的 MSA 通常是正确的，即使在高表达或选择性剪接转录产物存在的情况下也是如此（Marth et al. 1999），这样看来很有可能以此来克服虚假多态性的发生，因为 454 测序与经典的 Sanger 测序相比有较高的错误率可导致假的多态性。

POLYBAYES 介导的 SNP 挖掘的流程如图 26-2 所示，图中 454 EST 来自 Mo17 自交系玉米，锚定序列来自 B73 自交系。CROSS _ MATCH 提供的 MSAs 可以作为中心处理单元（CPU）密集，大规模的 454 序列组合最好预先利用 BLAST 进行分析使得每个 454 序列都能找到一个模板。随后 CROSS _ MATCH 在由模板和比对的 454 EST 序列的亚组合上运行。再由 POLYBAYES 评估 MSA 结果并找出 SNP。大规模序列分析不支持像插入或者替换这类的序列错误，但是这类错误很容易利用详细的序列比对分析检测出来。在撰写本章的时候，玉米基因组序列计划只差不到 20%就完成了。在临时期，高质量的（5000 bp 中只有 1 个不符）玉米基因组序列的汇集数据可以在 http://magi.plantgenomics.iastate.edu/下载（Emrich et al. 2004；Fu et al. 2005）。数据主要是来自 B73 基因组测量序列（GSS），是经过基因富集的基因组序列（Whitelaw et al. 2003）。

前期对玉米 454 EST 的 SNP 分析解读中使用 454 读数，每个碱基质量值为默认的

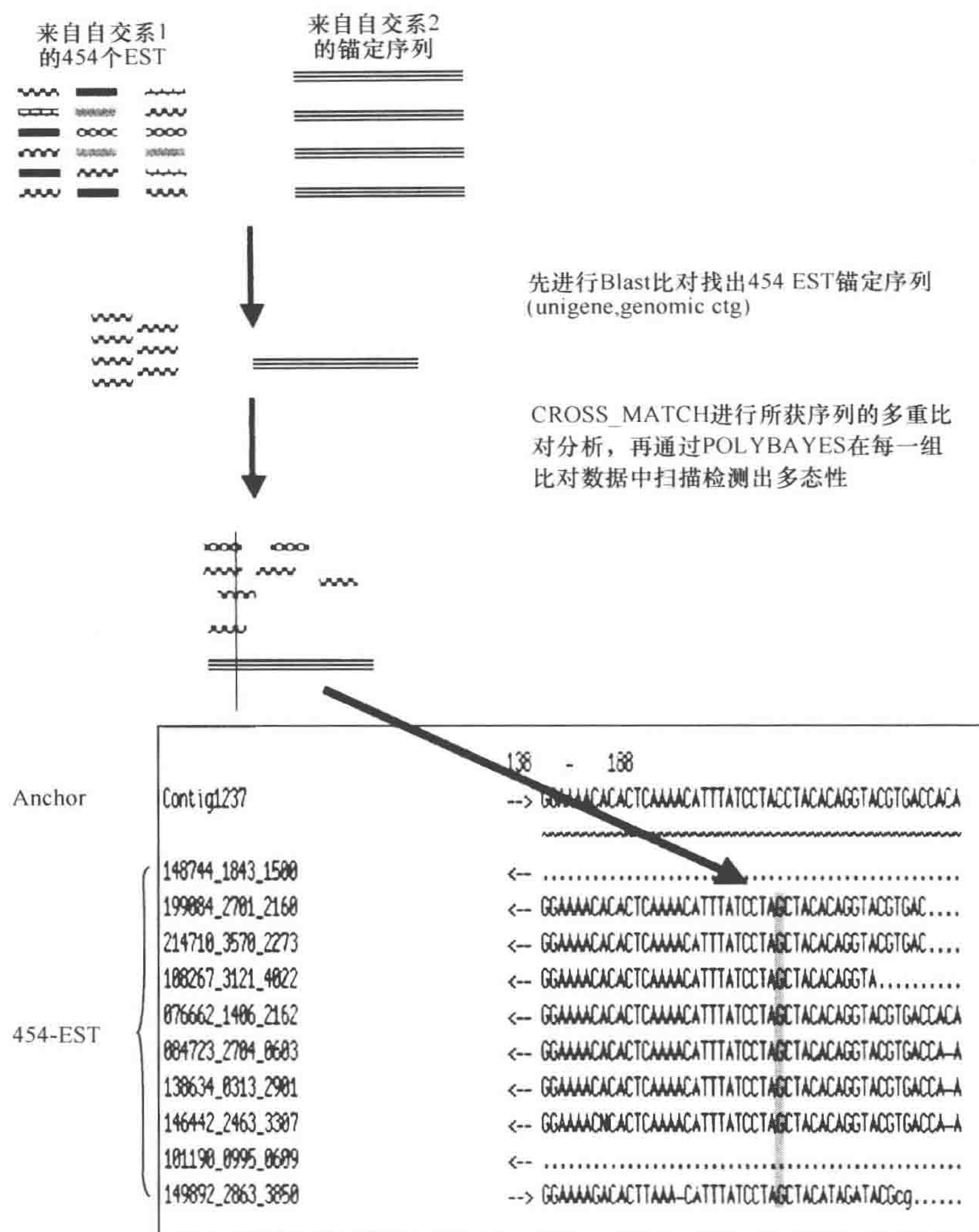



图 26-2　POLYBAYES 分析 454EST 序列检测 SNP 的流程图。EST 所关联的高质量 EST 或基因组序列才能作为锚定相用于模板驱动的 MSA。PLAYBAYES 扫描检测每个队列组合寻找 454EST 与模板之间可靠的多态性。上述例子中的 454 EST 来自 Mo17 玉米自交系，而锚定序列是来自自交系 B73。点 166 是 B73C 和 Mo17G 之间的多态性位点

18（W. B. Barbazuk，unpubl.）。这使错误率调整为约 1/65，这大大补偿了近来 454 测序中观察到的错误率（Margulies et al. 2005 ；Emrich et al. 2006）。尽管在 454 测序中的每个碱基读数都在 454 测序仪设定了质量标准，但这只有在经过被覆盖相同区域的独立测序验证后才是可靠的。

这里讨论的策略确定了 Mo17 和 B73 之间存在的超过 7000 个潜在 SNP。每个基因型覆盖范围影响每个 SNP 位点的可靠性，90％推测的 SNP 位点的取样在两个自交系中都要在 3 倍以上。一般来讲，高覆盖度的 SNPs 应该是有效的。在研究方案中，454 序

列设置默认质量值是依赖于碱基错误比率的经验评估值而不是利用454质量标准值。由此导致序列深度和相应等位基因比率极大地影响多态性检测，基于这样的观察结果，可通过检查每个多态性位点的统计结果来进行潜在SNP的筛查。POLYBAYES依赖于碱基质量和序列深度，因此准确的碱基质量值能够改进其性能。随着454测序技术成熟和454碱基质量参数精确度的增加，除了编制高质量的POLYBAYES使用文件之外，就将不再需要其他参数的修饰。进一步说，因为POLYBAYES的质量文件是可用的而不是放弃质量评价，所以POLYBAYES就非常适合于来自于混合平台序列的SNP检测（也就是说，将不具有质量分数的单一454读数与带有质量分数的454集合数据和Sanger数据结合起来）。

这里举一个例子，在接受一个给定的位点有多态性之前，需要确认B73和Mo17两个自交系在该位点上的每个核苷酸上都是单等位基因的（罕见的例外都几乎是由同一谬误引起的NIP；Emrich et al. 2004）。这是一个很大的制约，还会产生高的假阴性比率。设想一个位点，采样很好，并且在Mo17中是单等位基因的而且有10个B73的额外序列，它们中的9个都是一致的（但与Mo17等位基因不同），另外一个则不是。这个位点将被筛查出来，尽管唯一不一致的B73位点更像一个序列错误。所以，输出文件SNP的进一步分析可以被用来提高质量和精确性。

## 结论

因为水稻基因组序列目前已基本完成，进行水稻和玉米基因组序列的比对将以DNA为基础的分子标记定位在物理、遗传图谱，可以确定它们的潜在同线区域。这些区域提供了水稻-玉米同线性图谱的基础，在玉米基因组序列分析完成之前，同线性图谱能为玉米研究者提供一个便利的详细注释的基因组。下面是水稻-玉米同线性图谱的三个主要网络资源。

（1）Gramene(Jaiswal et al. 2006)，http://www.Gramene.org

（2）Arizona Genomics Computational Lab(AGCoL)，http:// www.agcol.Arizona.edu。

（3）TIGR，http://www.tigr.org/tdb/synteny/maize_IBMn/search_desc.html.。

玉米遗传图谱标记密度的逐渐增加，与禾本科植物基因组的同线性的结合，大大促进了基因图位克隆工程的实施（Bortiri et al. 2006）。

在两个玉米现代品种之间观察到的高度遗传变异很接近人类和黑猩猩之间遗传差异（Zhao et al. 2006）。这种很高的物种内遗传多样性奠定了玉米的表型多样性基础，使得玉米能适应多样性的环境。对遗传多样性范畴和机制的认识和理解对未来的植物育种是至关重要的。两个公共的玉米SNP筛查项目正在进行之中。威斯康星大学John Doebley负责的Panzea项目的目标是，检验过去筛选的SNP对玉米基因的分子多样性的影响，这个项目要进行大约5000个位点的SNP的验证，这些验证也被用来证明基因的确定性、多样化和选择的纯化。Wright等（2005）检验了774个来自玉米和玉米草随机基因的SNP序列，确定了一些筛选出来的候选基因，并得出了结论：有2%～4%的玉米基因经历过人工选择。接下来的研究（Yamasaki et al. 2005）验证了来自14个玉米

自交系的 1095 个基因序列，筛查那些可能参与驯化过程但不显示 DNA 序列多样性的基因。这个项目的信息可以应用后面的网站查询 http：//www. Panzea. org。

第二个公共玉米 SNP 搜索项目是由康奈尔大学 Ed Buckler 领导的，这个项目旨在基因组的 400 000 个位点中检测和评价 DNA 多态性，该研究将为今后的玉米的基因组关联作图提供特别的资源。

## 致谢

我们感谢 Mike Scanlon，现任及前任 Schnable 实验室人员 Mikio Nakazono、Dave Skibble 和 Marianne Smith 对本章实验研究方案所作出的贡献。本项研究由有多项竞标项目的资助，包括国家科学基金的植物基因组项目：P. S. S. （DBI-0321711 和 DBI-0321595）、W. B. B. （DBI-0501758）、the National Research Initiative of the USDA Co-operative State Research，Education、Extension Service P. S. S. （项目号 04-00913）、Hatch Act 和爱荷华州立基金 P. S. S. 及 ISU 的植物科学研究所和 Donald Danforth 植物科学中心分别对 P. S. S. 和 W. B. B. 的资助。

## 参考文献

Asano T., Masumura T., Kusano H., Kikuchi S., Kurita A., Shimada H., and Kadowaki K. 2002. Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: Toward comprehensive analysis of the genes expressed in the rice phloem. *Plant J.* **32:** 401–408.

Bainbridge M.N., Warren R.L., Hirst M., Romanuik T., Zeng T., Go A., Delaney A., Griffith M., Hickenbotham M., Magrini V., et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7:** 246.

Barbazuk W.B., Bedell J.A., and Rabinowicz P.D. 2005. Reduced representation sequencing: A success in maize and a promise for other plant genomes. *BioEssays* **27:** 839–848.

Batley J., Barker G., O'Sullivan H., Edwards K.J., and Edwards D. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* **132:** 84–91.

Baurle I. and Laux T. 2003. Apical meristems: The plant's fountain of youth. *BioEssays* **25:** 961–970.

Bortiri E., Jackson D., and Hake S. 2006. Advances in maize genomics: The emergence of positional cloning. *Curr. Opin. Plant Biol.* **9:** 164–171.

Bray N.J., Buckland P.R., Owen M.J., and O'Donovan M.C. 2003. *Cis*-acting variation in the expression of a high proportion of genes in human brain. *Hum. Genet.* **113:** 149–153.

Brendel V., Xing L., and Zhu W. 2004. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics* **20:** 1157–1169.

Burr B., Burr F.A., Thompson K.H., Albertson M.C., and Stuber C.W. 1988. Gene mapping with recombinant inbreds in maize. *Genetics* **118:** 519–526.

Ching A., Caldwell K.S., Jung M., Dolan M., Smith O.S., Tingey S., Morgante M., and Rafalski A.J. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet.* **3:** 19.

Coe E., Cone K., McMullen M., Chen S.S., Davis G., Gardiner J., Liscum E., Polacco M., Paterson A., Sanchez-Villeda H., et al. 2002. Access to the maize genome: An integrated physical and genetic map. *Plant Physiol.* **128:** 9–12.

Cone K.C., McMullen M.D., Bi I.V., Davis G.L., Yim Y.S., Gardiner J.M., Polacco M.L., Sanchez-Villeda H., Fang Z., Schroeder S.G., et al. 2002. Genetic, physical, and informatics resources for maize. On the road to an integrated map. *Plant Physiol.* **130:** 1598–1605.

Cowles C.R., Hirschhorn J.N., Altshuler D., and Lander E.S. 2002. Detection of regulatory variation in mouse genes. *Nat. Genet.* **32:** 432–437.

Davis G.L., McMullen M.D., Baysdorfer C., Musket T., Grant D., Staebell M., Xu G., Polacco M., Koster L., Melia-Hancock S., et al. 1999. A maize map standard with sequenced core markers, grass genome reference points and 932 expressed sequence tagged sites (ESTs) in a 1736-locus map. *Genetics* **152:** 1137–1172.

Dietrich C.R., Cui F., Packila M.L., Li J., Ashlock D.A., Nikolau B.J., and Schnable P.S. 2002. Maize *Mu* transposons are targeted to the 5′ untranslated region of the *gl8* gene and sequences flanking *Mu* target-site duplications exhibit nonrandom nucleotide composition throughout the genome. *Genetics* **160:** 697–716.

Dong Q., Lawrence C.J., Schlueter S.D., Wilkerson M.D., Kurtz S., Lushbough C., and Brendel V. 2005. Comparative plant genomics resources at PlantGDB. *Plant Physiol.* **139:** 610–618.

Emerson R., Beadle G., and Fraser A. 1935. A summary of linkage studies in maize. *Cornell Univ. Agric. Exp. Stn. Memoir* **180:** 1–83.

Emrich S.J., Barbazuk W.B., Li L., and Schnable P.S. 2006. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17:** 69–73.

Emrich S.J., Aluru S., Fu Y., Wen T.J., Narayanan M., Guo L., Ashlock D.A., and Schnable P.S. 2004. A strategy for assembling the maize (*Zea mays* L.) genome. *Bioinformatics* **20:** 140–147.

Fu Y., Emrich S.J., Guo L., Wen T.J., Ashlock D.A., Aluru S., and

Schnable P.S. 2005. Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. *Proc. Natl. Acad. Sci.* **102:** 12282–12287.

Fu Y., Wen T.J., Ronin Y.I., Chen H.D., Gou L., Mester D.I., Yang Y., Lee M., Korol A.B., Ashlock D.A., and Schnable P.S. 2006. Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. *Genetics* **174:** 1671–1683.

Gordon D., Abajian C., and Green P. 1998. *Consed:* A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Gowda M., Li H., Alessi J., Chen F., Pratt R., and Wang G. 2006. Robust analysis of 5′-transcript ends (5′-RATE): A novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Res.* **34:** e126.

Guo M., Rupe M.A., Zinselmeier C., Habben J., Bowen B.A., and Smith O.S. 2004. Allelic variation of gene expression in maize hybrids. *Plant Cell* **16:** 1707–1716.

Gut I.G. 2001. Automation in genotyping of single nucleotide polymorphisms. *Hum. Mutat.* **17:** 475–492.

Guyomarc'h S., Bertrand C., Delarue M., and Zhou D.X. 2005. Regulation of meristem activity by chromatin remodelling. *Trends Plant Sci.* **10:** 332–338.

Helentjaris T., Weber D.F., and Wright S. 1986. Use of monosomics to map cloned DNA fragments in maize. *Proc. Natl. Acad. Sci.* **83:** 6035–6039.

Hsia A.P., Wen T.J., Chen H.D., Liu Z., Yandeau-Nelson M.D., Wei Y., Guo L., and Schnable P.S. 2005. Temperature gradient capillary electrophoresis (TGCE)—A tool for the high-throughput discovery and mapping of SNPs and IDPs. *Theor. Appl. Genet.* **111:** 218–225.

Jaiswal P., Ni J., Yap I., Ware D., Spooner W., Youens-Clark K., Ren L., Liang C., Zhao W., Ratnapu K., et al. 2006. Gramene: A bird's eye view of cereal genomes. *Nucleic Acids Res.* **34**(Database issue)**:** D717–D723.

Jurinke C., Denissenko M.F., Oeth P., Ehrich M., van den Boom D., and Cantor C.R. 2005. A single nucleotide polymorphism based approach for the identification and characterization of gene expression modulation using MassARRAY. *Mutat. Res.* **573:** 83–95.

Kerk N.M., Ceserani T., Tausta S.L., Sussex I.M., and Nelson T.M. 2003. Laser capture microdissection of cells from plant tissues. *Plant Physiol.* **132:** 27–35.

Kristensen V.N., Kelefiotis D., Kristensen T., and Borresen-Dale A.L. 2001. High-throughput methods for detection of genetic variation. *BioTechniques* **30:** 318–326.

Kwok P.Y. 2001. Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum. Genet.* **2:** 235–258.

Lee M., Sharopova N., Beavis W.D., Grant D., Katt M., Blair D., and Hallauer A. 2002. Expanding the genetic map of maize with the intermated B73 x Mo17 *(IBM) population. Plant Mol. Biol.* **48:** 453–461.

Maher P.M., Chou H.H., Hahn E., Wen T.J., and Schnable P.S. 2006. GRAMA: Genetic mapping analysis of temperature gradient capillary electrophoresis data. *Theor. Appl. Genet.* **113:** 156–162.

Manaster C., Zheng W., Teuber M., Wachter S., Doring F., Schreiber S., and Hampe J. 2005. InSNP: A tool for automated detection and visualization of SNPs and InDels. *Hum. Mutat.* **26:** 11–19.

Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., Chen Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Marth G.T., Korf I., Yandell M.D., Yeh R.T., Gu Z., Zakeri H., Stitziel N.O., Hillier L., Kwok P.Y., and Gish W.R. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23:** 452–456.

Martienssen R.A., Rabinowicz P.D., O'Shaughnessy A., and McCombie W.R. 2004. Sequencing the maize genome. *Curr. Opin. Plant Biol.* **7:** 102–107.

Matsuoka Y., Vigouroux Y., Goodman M.M., Sanchez G.J., Buckler E., and Doebley J. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci.* **99:** 6080–6084.

Nickerson D.A., Tobe V.O., and Taylor S.L. 1997. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25:** 2745–2751.

Oeth P., Beaulieu M., Park C., Kosman D., del Mistro G., and Van den Boom D. 2005. iPLEX™ assay: Increased plexing efficiency and flexibility for Mass ARRAY® system through single base primer extension with mass-modified terminators. Sequenom Application Note No. 8876–006, Sequenom Inc., http://www.sequenom.com/Assets/pdfs/appnotes/8876-006.pdf

Pastinen T., Sladek R., Gurd S., Sammak A., Ge B., Lepage P., Lavergne K., Villeneuve A., Gaudin T., Brandstrom H., et al. 2004. A survey of genetic and epigenetic variation affecting human gene expression. *Physiol. Genomics* **16:** 184–193.

Piperno D.R. and Flannery K.V. 2001. The earliest archaeological maize (*Zea mays* L.) from highland Mexico: New accelerator mass spectrometry dates and their implications. *Proc. Natl. Acad. Sci.* **98:** 2101–2103.

Quackenbush J., Cho J., Lee D., Liang F., Holt I., Karamycheva S., Parvizi B., Pertea G., Sultana R., and White J. 2001. The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29:** 159–164.

Rabinowicz P.D. and Bennetzen J.L. 2006. The maize genome as a model for efficient sequence analysis of large plant genomes. *Curr. Opin. Plant Biol.* **9:** 149–156.

Rafalski J.A. 2002. Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci.* **162:** 329–333.

Rogers S.O. and Blendich A.J. 1985. Extraction of DNA from milligram amounts of fresh herbarium and mummified plant tissues. *Plant Mol. Biol.* **5:** 69–76.

Schnable P.S., Hochholdinger F., and Nakazono M. 2004. Global expression profiling applied to plant development. *Curr. Opin. Plant Biol.* **7:** 50–56.

Sharopova N., McMullen M.D., Schultz L., Schroeder S., Sanchez-Villeda H., Gardiner J., Bergstrom D., Houchins K., Melia-Hancock S., Musket T., et al. 2002. Development and mapping of SSR markers for maize. *Plant Mol. Biol.* **48:** 463–481.

Stupar R.M. and Springer N.M. 2006. *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics* **173:** 2199–2210.

Tenaillon M.I., Sawkins M.C., Long A.D., Gaut R.L., Doebley J.F., and Gaut B.S. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci.* **98:** 9161–9166.

Useche F.J., Gao G., Harafey M., and Rafalski A. 2001. High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Inform.* **12:** 194–203.

Wang J. and Huang X. 2005. A method for finding single-nucleotide polymorphisms with allele frequencies in sequences of deep coverage. *BMC Bioinformatics* **6:** 220.

Weckx S., Del-Favero J., Rademakers R., Claes L., Cruts M., De Jonghe P., Van Broeckhoven C., and De Rijk P. 2005. novoSNP, a novel computational tool for sequence variation discovery. *Genome Res.* **15:** 436–442.

Wei H., Fu Y., and Arora R. 2005. Intron-flanking EST-PCR markers: From genetic marker development to gene structure analysis in *Rhododendron. Theor. Appl. Genet.* **111:** 1347–1356.

Whitelaw C.A., Barbazuk W.B., Pertea G., Chan A.P., Cheung F., Lee Y., Zheng L., van Heeringen S., Karamycheva S., Bennetzen J.L., et al. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302:** 2118–2120.

Wright S.I., Bi I.V., Schroeder S.G., Yamasaki M., Doebley J.F., McMullen M.D., and Gaut B.S. 2005. The effects of artificial selection on the maize genome. *Science* **308:** 1310–1314.

Wu T.D. and Watanabe C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21:** 1859–1875.

Yamasaki M., Tenaillon M.I., Bi I.V., Schroeder S.G., Sanchez-Villeda H., Doebley J.F., Gaut B.S., and McMullen M.D. 2005.

A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17:** 2859–2872.

Zhang J., Wheeler D.A., Yakub I., Wei S., Sood R., Rowe W., Liu P.P., Gibbs R.A., and Buetow K.H. 2005. SNPdetector: A software tool for sensitive and accurate SNP detection. *PLoS Comput. Biol.* **1:** e53.

Zhao W., Canaran P., Jurkuta R., Fulton T., Glaubitz J., Buckler E., Doebley J., Gaut B., Goodman M., Holland J., et al. 2006. Panzea: A database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.* **34**(Database issue): D752–D757.

# 互联网资源

**http://www.agcol.arizona.edu** Arizona Genomics Computational Laboratory

**http://www.complex.iastate.edu/download/GRAMA/index.html** GRAMA (Genetic Recombinant Analysis and Mapping Assistant), Complex Computation Laboratory, Iowa State University.

**http://www.gramene.org** GRAMENE: A resource for comparative grass genomics

**http://www.tigr.org/tdb/synteny/maize_IBMn/search_desc.shtml** TIGR Rice Genome Annotation, Maize Genetic Markers Mapped to Rice Pseudomolecules

# 27 水　　稻

Hei Leung,[1] Kenneth L. McNally,[2] and David Mackill[1]

*[1]Plant Breeding, Genetics, and Biotechnology Division; [2]T.T. Chang Genetic Resources Center, International Rice Research Institute, Manila, Philippines*



## 简介

水稻（*oryza sativa*）是最古老的禾谷类作物之一，据报道水稻的种植史可追溯到 7000 年前（Wanabe 1997）。水稻种植在广阔的地理区域，几乎遍及世界上的所有大陆，最大的集中种植区在亚洲（Khush 1997）。水稻可以在许多不同的生态区域，包括干旱的丘陵地和可以控制灌溉的洼地以及深水环境中种植。这些生态系统很大程度上反映可利用的水资源，并影响着被驯化以后的水稻的适应性和自然演化。通过长期种植栽培中的自然选择和人工选择使水稻获得了令人惊叹的遗传多样性。

栽培水稻是二倍体，具有 12 对染色体（$2n=24$）。一些人认为被驯化栽培的水稻有两个亚种：*indica* 和 *japonica*，在水稻生长区域被广泛种植；另一些人认为这两个“亚种”只是生态型或分类学来源不明的栽培种。至于生产，indica 水稻是种植在热带和亚热带地区的优势类型，提供世界水稻产量的 80%（Mackill 1995）。而 japonica 水稻是

在温带地区种植的主要类型。水稻种植主要受消费者的喜好和环境需求的影响，同时，这两个变种的某些特征也是相互重叠的，它们的基因池也经常发生混杂。另一个被驯化的水稻二倍体种是 *Oryza glaberrima*，最早种植在非洲。除了栽培水稻之外，野生型稻及其近缘种属的多样性包括二倍体和四倍体种。

水稻中固有的丰富遗传多样性不仅为遗传改良提供了基础，而且也是研究作物进化和栽培驯化史的一个优异的资源。在所有作物物种中，水稻幸运地被赋予异常丰富的遗传资源，使之可以支持先进的遗传学技术和工具的一种遗传系统的模式。水稻是已知的第一个完成基因组测序的农作物。国际基因组测序的共同努力已得到 indica 和 japonica 两个水稻变种的基因组全部信息。伴随着其他植物物种基因组序列信息的扩展，提供了一个深入认识、开发和利用水稻遗传资源的新模式（Leung and Au 2004）。

在新模式下，水稻遗传学知识能通过比较基因组学分析进行物种间的整合，也能够加速对基因功能的揭示。进一步说，基因组范围的分析具有揭示全新遗传途径的巨大潜力，也为迎接揭示可预见的和不可预见的遗传功能的挑战提供了新的机遇。实现这样的潜能，不论天然还是诱导的丰富遗传多样性资源必须被建立或者是可以被利用的。国家的或国际的基因库中保存的种质资源需要被进一步的鉴定，确定其遗传学特性。否则的话，巨大的遗传多样性储备资源将一直不能被利用。因而，保守性、世代性和遗传多样性特征及基因功能的发现是一个相互依赖的活动，需要整合研究才能取得预期的成果。这一章将要讲述利用遗传资源探索水稻遗传和功能多样性的最新进展，以及如何运用水稻的遗传资源。我们也讨论如何更好地利用基因组信息其他资源。我们希望这一简短的叙述能为令人振奋水稻遗传学研究提供概况，特别是那些最终将对水稻的改良产生实质的影响的领域。

## 水稻的遗传变异性

国际水稻研究所（IRRI）的国际水稻基因数据采集中心（IRGC）有世界上最大的水稻种质收藏，包括驯化的水稻及其野生近缘种属。目前，这里有 102 000 份来自亚洲栽培水稻 *O. sativa*，1650 份来自非洲栽培水稻 *O. glaberrima* 的变种，以及 4510 份来自 22 个野生近缘种的变种或亚种。表 27-1 总结了 IRRI 基因库中保存的水稻种质多样性资源宝藏。该收藏提供了极其深厚基因库，其中大部分的功能和特征还尚未弄清楚，因此针对这一多样性库的认识、开发和应用研究的高效技术是非常需要的。

亚洲栽培水稻存在着大量的变异类型，基于同功酶分析大致可分为 6 个组。在 1980 年代初期，Second（1982，1985）通过同工酶等位基因分析的研究提出假说，认为 indica 和 japonicia 两个亚种是 *O. sativa* 水稻经过两个独立驯化过程演化而来的。最近，indica 和 japonicia 分别独立驯化的假说有了新的证据，这一新证据的获得是基于 SINE（short interspersed nuclear element）模式分析（Cheng et al. 2003）、长末端重复序列的（LTR）反转座子的转座史（Vitte et al. 2004）、SSR 的群体结构分析（Garris et al. 2005），以及核基因内含子的系统演化分析（Zhu and Ge 2005）等。作为两个最重要的水稻类型是分别驯化来的，这对关联遗传学的应用是非常重要，因为群

体结构是必须考虑的因素。图 27-1 显示出，通过 SSR 基因分型获得的驯化水稻主要类型和他们假定的起源种之间的全部关联。该图主要是基于 Generation Challenge programme project（www.generationcp.org）的工作，该项目由 IRRI、CIRAD、CIAT、EMBRAPA、WARDA 和康奈尔大学合作开展（K. McNally et al.，unpubl.），类似的工作 Garris 等（2005）也有过报道。该图的结构反映同工酶多样性分析的 5 个组群，即 indica、aus、aromatic 和 japonica，后者被细分为热带和亚热带两种类型。野生型 AA 基因组物种位于 aus/indica 和 aromatic/japonica 组群之间，源于两个或多个主要组群个体的等位基因比例显示出混合型。

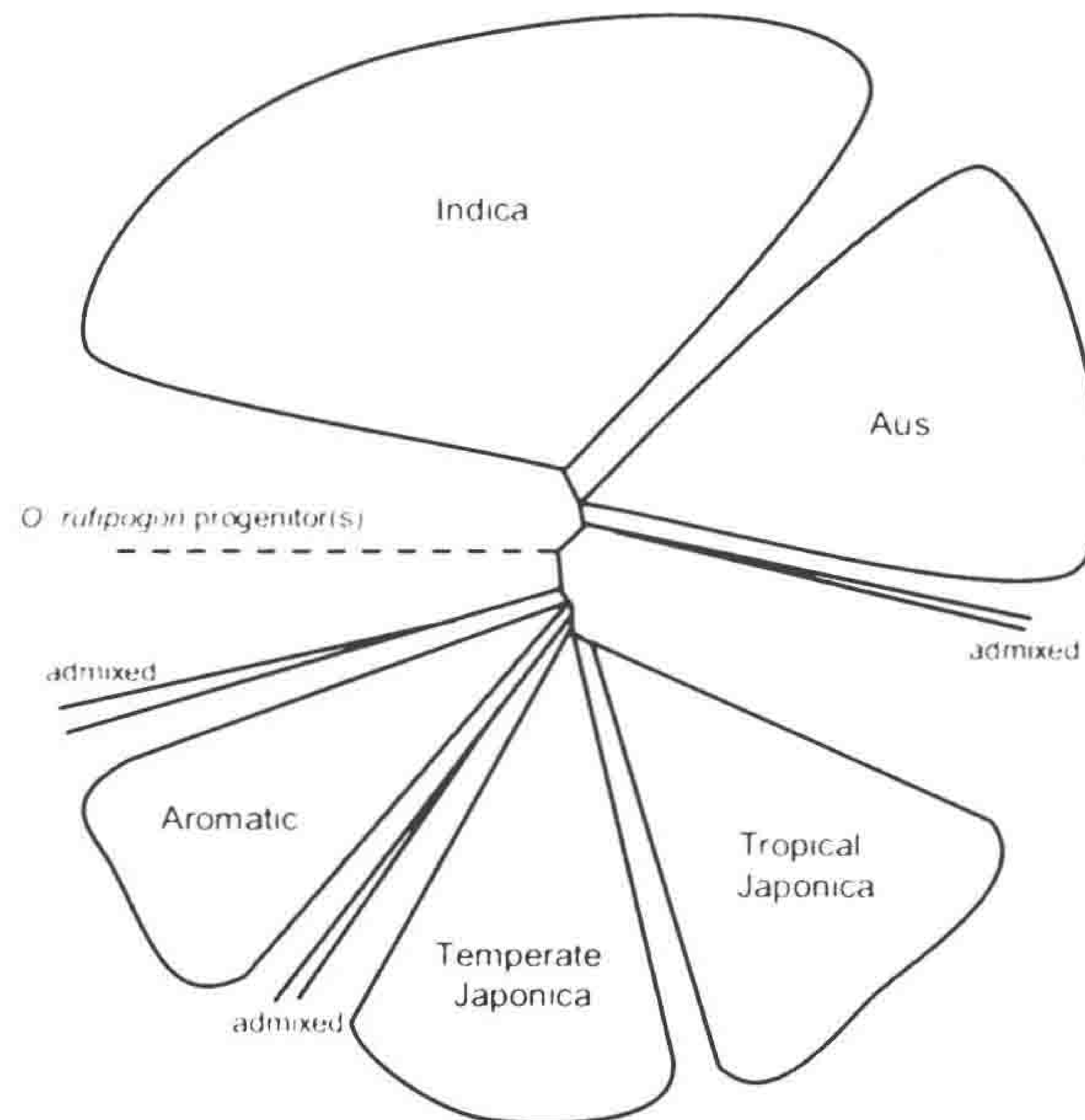



图 27-1　水稻（O. satia）主要品系分类组之间的遗传多样性关系图示，该图基于 Generation Challenge programm project（www.generationcp.org）SSR 基因分型工作，该项目由 IRRI、CIRAD、CIAT、EMBRAPA、WARDA 和康奈尔大学合作开展（K. McNally et al.，unpubl.），类似的工作 Garris 等（2005）也有过报道。水稻野生的起源祖先用虚线表示，混合型所标示的引线是来自两个或两个以上主要组分的等位基因比率

现代分类学发现野生型水稻 *Oryza* 有 22 个物种，这些物种绝大部分是二倍体，带有 5 个异源四倍体类型（表 27-1）。这些物种有 10 种不同的基因组类型。按照 Khush（1997）的分配：AA 基因组（包括 Sativa 集群由栽培种 *O. sativa* 和 *glaberrima*，野生种 *O. rufipogon*、*nivara*、*glumaepatula*、*meridionalis barthii* 和 *longistaminata*）；BB、CC、BBCC 和 CCDD 基因组（officinalis 集群有 *O. alta*、*punctata*、*eichingerri*、*minuta*、*officinalis*、*rhizomatis*、*grandiglumis* 和 *latifolia*）；EE 基因组（*O. australiensis*）；FF 基因组（*O. brachyantha*）；GG 基因组（Meyeriana 集群包括 *O. granulata*、*meyerian* 和 *neocalidonica*）；HHJJ 基因组（Ridleyanae 集群包括 *O. longiglumis* 和 *ridleyi*）；HHKK 基因组（*O. schlechteri*）。图 27-2 显示了带有 AA 基因

**表 27-1 人工栽培和野生水稻种在国际水稻基因库中的基因组**

| 基因 | 种属[a] | 分布区域[b] | 基因位置[c] |
|---|---|---|---|
| AA | O. sativa | 全球 | |
| AA | O. glaberrima | 非洲西部 | 102 533 |
| AA | O. barthii | 非洲 | 1 650 |
| AA | O. nivara | 亚洲热带和亚热带地区 | 211 |
| AA | O. rufipogon | 亚洲热带和亚热带地区、大洋洲的 | 1 260 |
| | | 热带地区 | 1 048 |
| | | 非洲 | 218 |
| AA | O. longistaminata | 美洲中部和南部 | |
| AA | O. glumaepatula | 大洋洲的热带地区 | 54 |
| AA | O. meridionalis | 全球 | 55 |
| AA | hybrid or uncertain classifica- | 非洲 | 1 002 |
| BB,BBCC | tion | | 75 |
| BBCC | O. punctata | 菲律宾、巴布亚新几内亚 | 63 |
| CC | O. minuta | 亚洲热带和亚热带地区、大洋洲热带 | 279 |
| | O. officinalis | 地区 | 20 |
| | | 斯里兰卡 | 24 |
| CC | O. rhizomatis | 亚洲南部、非洲东部 | 7 |
| CC | O. eichingeri | 美洲中部和南部 | 58 |
| CCDD | O. alta | 美洲中部和南部 | 10 |
| CCDD | O. latifolia | 美洲中部和南部 | 36 |
| CCDD | O. grandiglumis | 大洋洲热带地区 | 17 |
| EE | O. australiensis | 非洲 | 24 |
| FF | O. brachyantha | 亚洲南部、亚洲东南部 | 10 |
| GG | O. granulata | 亚洲东南部 | 1 |
| GG | O. meyeriana | 新喀里多尼亚 | 6 |
| GG | O. neocaledonica | 印度尼西亚、巴布亚新几内亚 | |
| HHJJ | O. longiglumis | | |
| | | 亚洲南部 | 14 |
| HHJJ | O. ridleyi | 巴布亚新几内亚 | 1 |
| HHKK | O. schlechteri | | |

a. 来自于 Leung 和 An(2004),及 Lu(1999)表格的修改。

b. Brar 和 Khush(1997)

c. 信息来自 IRGC 数据库,2006 年 10 月 25 日的网站资料:www.irgcis.irri.org:81/grc/irgcishome.html。

组的 *Oryza* 物种的表型多样性,该物种在水稻育种中经常使用。

在 OMAP(Oryza Map Alignment Project,www.omap.org)项目中,13 个代表性的野生型水稻变种的基因组被用来构建 BAC 文库(Ammiraju et al. 2006;Wing et al. 2007)。克隆的 BAC 末端序列和指纹图谱分析使之锚定到 Nipponbare 基因组及其重叠群。根据不同基因组的物理图谱,选定特异部位的相关基因的克隆可被筛选出来进行深入的研究。这些文库和物理图谱提供了深入认识水稻基因组之间的遗传距离和多样性的有力工具。*O. sativa* 水稻和野生近缘种的广泛杂交种通过基因渗入,使许多响应

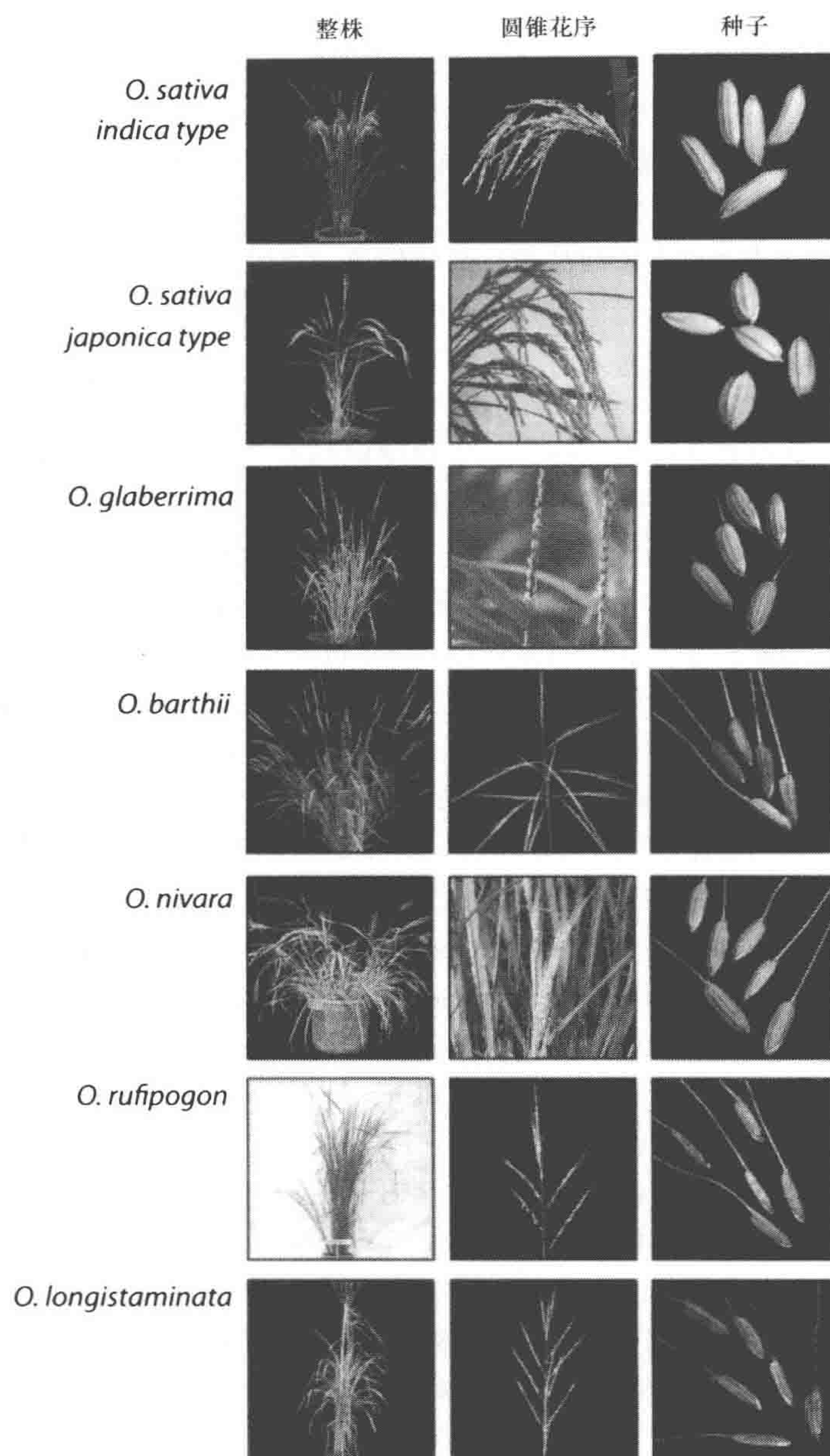



图 27-2　带有 AA 基因组水稻（*Oryza*）典型种的表型多样性示意图。代表种在整株、圆锥花序和种子三个发育时期北部获得的一些显著特征。因为水稻 AA 基因组种是互交品系，因此改基因组类型的品种在水稻品种改良种被广泛应用

生物和非生物胁迫的有益性状导入了现代水稻栽培种（Brar and Khush 1997）。这些包括来自 *O. nivara* 的草丛短缩病毒抗性（Grassy stunt virus）基因，来自 *O. longistaminata* 的细菌性枯萎病抗性，来自 *O. rufipogon* 的水稻东格鲁病抗性和来自 *O. australiensis* 的褐飞虱抗性基因等。

除国际水稻基因数据库而外，国际水稻研究所（IRRI）也拥有水稻遗传进化的国际网站（INGER），主要包括筛选目标环境下的多种生物学特征。该机构拥有从世界范围内的育种项目中收集到的 40 000 份水稻品种的储藏量。此外，所有 IRRI 选育品种和许多非 IRRI 培育的现代水稻品种的育种系谱，其中也有许多是来自传统品种的交配体系，这些资料都保存在国际水稻信息系统（www.iris.irri.org）；除了家谱数据，IRIS 还保存有来自 IRRI 和 INGER 的品种特征的信息。

## 水稻物理图谱

1997 年，在新加坡举行的植物分子生物学国际会议上第一次讨论了关于水稻全基因组计划。1998 年国际水稻基因组测序工程（IRGSP）开始启动，来自 10 个国家和地区的众多研究机构和实验室承担这项工作。在不到 10 年的时间里，科学界获得至少两个水稻品种的全基因组序列，即日本稻品种 Nipponbear 和印度稻品种 93-11。两个水稻品种基因组序列草图是由北京基因组中心（Yu et al. 2002）和 Syngenta 公司（Goff et al. 2002）采用全基因组鸟枪测序法最先绘制的。实施这种方案，基因组克隆被随机测序并进行信息整合。IRGSP 则采用逐一克隆测序的方法，利用该策略 BAC 文库的大量插入克隆第一次排列出来产生一个毗邻的物理图谱。然后，每一个 BAC 文库进行小插入片段的测序。碱基序列图需要付出了更多的时间和精力，但这被认为是 DNA 序列的染色体定位和核苷酸确定的黄金标准。2005 年 Nipponbare 的全部 12 条染色体的最后图谱被公布（IRGSP 2005）。在全部基因组序列公布之前，某些单条染色体完整序列已被公布出来（Feng et al. 2002；Sasaki et al. 2002；水稻染色体 10 Sequence Consortium 2003；水稻染色体 3 Sequencing Consortium 2005）。早期的单个染色体序列的公开释放使整个科学界活跃地运用基因组序列信息进行杂交种的分析，促进了 SNP 和插入、缺失作为选择性标记的应用（Feltus et al. 2004；Shen et al. 2004）。

IRGSP 确定了 Nipponbare 品系的 370 Mb 的核酸序列，覆盖了整个基因组 398 Mb 的 95%，几乎覆盖了全部常染色体区。考虑可转录原件相关序列，着丝粒区和端粒区，大约 60%的基因组序列（240 Mb）是由单一序列和低拷贝序列组成。总共发现 37 544 个蛋白质编码基因（不包括转录元件相关基因）。这些结果提供了一个蛋白编码基因保守性估价，这些已经通过基因序列标签 EST 或 cDNA 全长序列得到实验的证实。这样的话，基因的密度是平均每 9.9 kb 一个（IRGSP 2005；Matsumoto et al. 2007）。所有的基因组序列都可以通过水稻注释数据库（http://rapdb.lab.ac.jp）和基因组研究（TIGR）网站（www.tigr.org/tdb/e2k1/osa1/）搜寻和公开使用。考虑到许多预期的基因未被表达数据库支持，这些数据的功能元件数目可能会更大。同时还存在一个非编码 RNA 基因的潜在序列储备，这些基因的功能至今仍知之甚少（Sunkar et al. 2005）。

## 基因组 SNP 的检测

Nipponbare 和 93-11 品种基因组序列的公布使开展水稻 japonica 和 indica 两种类型之间的基因组 SNP 分析成为可能。利用这两个基因组，Feltus 等（2004）在去除多拷贝和低质量序列之后，确定出大约 400 000 个 SNP 或插入/缺失片段（indel 候选位点，并通过随机选择性测序使 109 位点得到验证。其中有 87 个含有 SNP 的位点在数据库得到匹配，预测结果给出了约 80%的精确性。另外的分析中，通过 12Mb 的 indica Kasalath 与 japonica Nipponbare 序列对比，IRGSP 预测在 indica 和 japonica 之间有 80 127 个多态性位点。按照这一频率，整个基因组之间将产生 260 万 SNP，相当于每 1000bp 有 6 个。Yu 等（2005）曾预测在编码序列中每 1000 bp 中有 3 个 SNP，而在转录元件中可高达每千个碱基对有 28 个 SNP。

Shen 等（2004）在 Nipponbare 和 93-11 进行 SNP 和 indel 检测分析，发现每 268 bp 有 1 个 SNP，每 953 bp 有 1 个 indel。也就是说，在这两个栽培品种中，SNP 多样性的出现频率是 indel 的多样性的 3 倍。然而，SNPs 在一个特定的染色体中不是均匀分布的。在 4、5、8 和 10 号染色体的着丝粒区的多态性密度是很低的，但是在 7 号染色体的着丝粒区多态性密度是很高的。不管是那种分析方法，非常清楚的是在 japonica 和 indica 两种水稻的典型品种之间确实地存在着大量的多态性序列。

基因组中存在丰富的 SNP，使得建立 SNP 的模式和表型之间的联系成为可能。这里有一个关键问题是要弄清楚这些 SNP 在代表水稻不同基因池的多个基因组中是如何分布的。全基因组 SNP 的检测和鉴定可提供一些物理距离相近的 SNPs 的信息（SNP 单倍型），这些信息能反映出进化和功能上相互关系，也能提供一个其他相关的多态性（包括生化、代谢和生理方面）与表型的锚定点。从实际应用而言，一个多品种的 SNP 数据库将比一两个水稻基因组序列有更加深远的意义，因为只有通过对比分析，我们才能将不同品种间的等位基因差异与转录水平和蛋白合成的改变、代谢上细微但关键的变化以及最终的表型变化联系起来。

### 水稻 20 个不同品系基因组 SNP 的对比分析

意识到从不同的基因型中分析获得 SNP 数据的可能性，水稻功能基因组协会与 IRRI 一道共同努力，承担起对基因组 SNP 的检测任务，该检测通过高密度寡聚核苷酸芯片的 DNA-DNA 杂交进行水稻不同品种的测序分析来实现（McNally et al. 2006）。该工程是由 Perlegen Science 提供用于对照模板确定变异序列的嵌合芯片（参考基因组），用这种重复测序的方法，整个基因组中长度大于 60 bp 的非重复区与用 25 个碱基的寡核苷酸组合拼贴通过两条链的单个碱基移动而得到，在中间区域是 4 倍的降低。因此，芯片上的 8 个寡单核苷酸探针鉴定非重复区域中的一个位点。

目前，25 个不同的品种（或地方种）已收集起来进行扩种和纯化。其中已经有 20 个品系用在了 SNP 检测项目中。这些品系已经完成了单一种子的扩繁，获得足够用于基因分型的数量。另外，为准备构建重组自交系（RIL）和其他群体，这些品系的半双

重杂交计划的 $F_1$ 代互交业已完成。

利用 McGill 大学的 Thomas 研究所建立的管道体系，将第四次公布的日本稻 Nippnbare 高通量 BAC 文库序列的重复序列遮盖。遮盖后，通过每个片段的 BLAST 除了自身之外没有其他明显相同序列的严格标准分析，日本稻 Nipponbare 基因组中有 95 Mb 的序列是单拷贝的。同样的程序也被在印度稻 93-11 鸟枪测序法获得的基因组序列分析（北京基因组中心）。两个基因组序列用 Mummer/Nucumer 工具（TIGR）进行平行相互比对。通过与 TIGR 基因模型进行比较，被遮盖的序列确定在基因组的 100 Mb，这些序列将被整合到芯片中。整合的序列被分散到整个基因组区域，使基因组的每 100 kb至少在芯片中出现一次。20 个不同品系分别与这些芯片杂交将能检测出覆盖整个基因组范围内的 SNP。

利用这些芯片完成了一个引领性的实验，3 号染色体长臂的 684 kb 区域检测出 378 kb的单一序列。20 个全部品系与这类芯片杂交，总共检测到 2132 个 SNP 位点(平均 200bp 有 1 个 SNP)。由几个日本稻品系的比对，在 379kb 的区域只检测到少于 10 的 SNP。现在还不太清楚，这些品系中多样性水平低的特点是仅仅在这个 379 kb 的区域还是整个基因组；当全基因组数据完成时这个问题就有答案了。最新的研究(Tang et al.　2006）显示在印度和日本稻类型上的差异可能是由于两个基因组的不同造成的。他们发现两种类型的水稻基因组之间至少 6%的区域有不寻常的差异，其他相关区域显示出较低的差异。这些差异的显著区域应该能反映了驯化的历史。

## SNP 单倍型

SNP 与高度的连锁不平衡（LD）有很高的相关性。染色体上等位基因的特定组合决定其单倍型。比较杂交个体间单倍型和高连锁不平衡在减数分裂过程中的分离，这样就可以识别出与微小片段重组的相关片段或单倍型区段。由一个等位基因单倍型区段可以预测该区段的其他等位基因，每一区段一个或几个等位基因的基因分型就可以充分确立特定区域的单倍型。这样预测的 SNP 称为 tag-SNP。因此，全基因组扫描可以通过限定单倍型区域的 tag-SNP 的收集完成基因分型。SNP 预测数据库的建立将有助于确定 tag-SNP 的数量，揭示整个水稻基因库中所有的单倍型变异。

高密度低聚物重新测序方法，已经在人和小鼠的基因组中检测到 SNP 单倍型(Frazer et al.　2004；Hinds et al.　2005)。在人类中，通过 24 个代表性个体的微阵列为基础的重新测序技术确定有 160 万个 SNP 在不同家系的个体中是共有的（Hinds et al.　2005)。当采用这些 SNP 资料进行一个 71 人样本组的基因分型时，检测到高于 73%的共有变异。一个国际联盟已经绘出人类基因组的单倍型图谱（International HapMap Consortium　2005)。作为一期图谱，有 100 万个 SNP 被确定，基于来自 4 个(基因组中大约每 5kb 有 1 个 SNP）人群的 269 人 SNP 基因分型研究。关于小鼠，Frazer 等（2004）利用 13 个典型的近交系和野生近交系完成了总共 4.6 Mb 跨越 5 个基因组区域 SNP 搜索。这一分析揭示了一些从 12～608 kb 单倍型区域，这些区域能简单地直接划分小鼠 13 个典型近交系的亲缘关系。这项研究小鼠中 SNP 被尽数搜索出来，在 12 个小鼠近交系中发现 97%的祖先共有的变异的遗留。

已有两项叙述水稻连锁不平衡范畴的研究发表。Garris 等（2003）研究了 5 号染色体短臂上的 xa5 位点周围的单倍型多样性和连锁不平衡，发现间隔大致为 100 kb 的区域有明显的连锁不平衡。Olsen 等（2006）发现在 6 号染色体 Waxy 位点延长到 250kb，另有其他的多样性检测表明在人工驯化过程中这个位点被选择性清理过。因此，水稻这个单倍型区域的长度为 100～250kb。为在水稻深入这些连锁不平衡研究，CIRAD 和 IRRI 测定了决定水稻差异的单倍型结构的范围。适于基因组扫描的一组共 1536 个 SNP 被用于进行 900 个不同类型水稻的基因分型，该工作可以用来预测覆盖 indica/japonica 的多样性和预期它们之间可能的天然杂种。

对整个基因组范围的 SNP 分析，将会得出不同水稻品种共有的 SNP 变异的精细连锁不平衡图谱。这将阐明旱季矮秆稻和芳香型水稻变种的起源与日本和印度水稻的关系。在一定性状范围内，结合表型数据的 SNP 分析将用于决定水稻品系的位点的检测和鉴定。这可能揭示性状和 SNP 的相关性，使对农业有益的基因的连锁不平衡为基础的绘图成为可能，并加速在其他近交种的应用。但不管怎样，重要的是要注意水稻有多种遗传杂交选择，没有必要建立非常高分辨率的连锁不平衡图谱。可以明确地讲，基因型和表型之间的关系能够通过其他数据资料或经济有效的实验方法更好地鉴定出来。

## 等位基因变异的识别

由于连锁不平衡结构，全面揭示带有 SNP 单倍型区域的水稻种质的自然变异成为可能，这些单倍型区域是保守的或者在水稻驯化的年代经过选择的。如果 SNP 单倍型和表现型之间的关联是稳定的，就增加了在获选基因中发现具有特殊功能的有用等位基因的可能性。进一步说，通过 SNP 单倍型分型建立的遗传群体，可为在水稻基因池中有效地探索等位基因多样性奠定了基础。新的改进技术在目前是可用的。例如，一种称为 TILLING（基因组的定向诱导损伤）的技术，是从用于检测点突变的反向遗传学技术发展而来的（McCallum et al. 2000）。另一种是 Eco TILLING，这项技术主要适于检测个体基因组位点的自然变异（Comai et al. 2004）。

第一步就是，从水稻的种质资源中选出栽培水稻 *O. sativa* 的变种/品种共 1536 个亚组，作为微核心组合。核心组合代表亚洲栽培水稻种质的地理分布区域，及其遗传多样性的分布范围。非洲稻栽培种 *O. glaberrima* 是以 190 变种作为代表。除驯化的水稻而外，我们也利用水稻的 93 个野生近缘变种，代表野生水稻 Oryza 属的所有基因组类型（AA、BB、CC、BBCC、CCDD、EE、FF、GG、HHJJ、HHKK）。关于 AA 基因组类型，我们利用 95 个有代表性变种阐述了 *Oryza* 相关近缘种之间的系统关系。

我们通过对 *O. sativa* 的微核心组合和 *O. glabberima* 的 10 个基因，和野生型 *Oryza* 水稻 93 个变种亚组合的 5 个基因的初步检测结果分析，证明 EcoTILLING 技术是水稻种质资源收藏品种特征分析的一种有效方法。第一，检出的 SNPs 频率与预期的分析结果是一致的。第二，针对预测保守序列的引物设计是以日本水稻 Nippobare 序列功能为基础（如 DREB2 的编码区），这些序列功能可涵盖所有 AA 基因组和大多数该基因组以外的基因组。尽管一对引物序列跨基因组工作的有效性取决于该序列的保守程度，

但是评价基因组之间 SNP 和 In/del 等位基因变异的能力为我们创造了新的机遇。第三，用最简单的琼脂糖电泳进行 EcoTILLING 分析，可用于栽培水稻和大多数含 AA 基因组的 *Oryza* 种的水稻分析，并能提供一种快速、经济有效的检测和描述变异特征的手段（Raghavan et al. 2007）。

## 作图及确定基因功能的遗传资源

特异性遗传谱系和作图群体是研究基因功能的必要条件，如果我们想利用水稻的全部基因序列信息，我们必须解决好这个先决条件的建立。关于特殊遗传谱系有两种途径可以考虑，第一种是诱变，可以利用选定的基因型（经常是通过表现型的选择得到的）去诱导产生位点变异。水稻的诱变方法发展得很好，可以利用内源的或外源的转座子，插入性载体及化学诱变或物理射线辐照处理等。诱导突变优越性在于其遗传背景几乎是一致的（不包括转化诱导的体细胞变异），这种突变体和野生型可以看作是遗传同源系，使得它们的比较分析变得十分简单容易。目前，尽管距离突变体饱和覆盖整个基因组还有差距，但是一个数量相当可观的突变体库已经可以使用了（Hirochika et al. 2004）。不管怎样，诱导突变还是受到被选定基因型位点的等位基因多样性（同一位点两个可替代的等位基因）的限制。

第二种特异性遗传谱系可以从物种中已经存在的自然变异捕获。通过研究不同遗传背景下杂交的遗传因子的研究并创建新的遗传重组子。尽管通过不同的遗传研究或育种程序已经获得了许多的作图群体，但是能进行系统维护和公开散布共享使用的重组群体依然很少。因此，合理设计的可用于基因功能分析的重组群体仍是一个很大的需求。

### 诱导变异

由于国际上众多国家共同努力创制用于功能基因组研究的突变体库，水稻获得了特别丰富的诱导的遗传变异。与任意一种农作物相比，水稻诱导突变体库可能是规模最大的（Hirochika et al. 2004），这其中包括转座子或 T-DNA 插入导致丧失或超表达基因功能的株系。2006 年 11 月，Oraygene DB（http：//oraygenedb. cirad. fr）汇编水稻基因组的带有已知两翼序列标签的 14 万个插入序列（表 27-2）。通过这些插入的突变体库，有 50％的可能性发现任意一个基因的突变标签（A. Pereiera，pers. comm.）。带有不同整合特性的多重插入载体和转座子的应用有助于全基因组的覆盖。但是，用插入诱变来达到基因组饱和性突变可能会很难，因为有多基因或基因组区域可能是不易被插入序列整合的。要使突变达到基因组饱和，化学和辐射诱变是对插入突变的一个很好的补充。Bhat 等（2007）对水稻化学和辐射诱变现状做了详细的总结。与转化和组织培养介导的突变方法不同，化学和辐射诱变能作用于任何一种基因型。这些突变体的非转基因特性使它们在广泛分布上有特殊的应用，因此可以对其在与农业相应环境的生物特征作出评价。

表 27-2 国际研究机构所获得带有 FSTs 的水稻插入突变体情况

| 研究机构 | 联系方式 | 载体 | 序列号 |
|---|---|---|---|
| CSIRO | narayana. upadhyaya@csiro. au | T-DNA | 787 |
| CIRAD-INRA-IRD-CNRS,Genoplant | emmanuel. guiderdoni@cirad. fr | T-DNA | 7 480 |
| 美国国立农业生物科学研究协会 | hirohiko@nias. affrc. go. jp | *Tos17* | 18 024 |
| GerealGene Tags,European Union | andy. pereira@wur. nl | *Ds* | 1 380 |
| Gyeongsang National University(庆尚大学) | cdhan@nongae. gsnu. ac. kr | *Ds* | 1 040 |
| Postech | genean@postech. ac. kr | T-DNA | 80 259 |
| 国家植物基因研究中心(武汉) | swang@mail. hzau. edu. cn | T-DNA | 15 727 |
| 新加坡国立大学 | sri@tll. org. sg | *Ds* | 1 469 |
| 台湾水稻插入突变体计划 | bohsing@gate. sinica. edu. tw | T-DNA | 7 053 |
| 加利福尼亚大学(戴维斯分校) | sundar@ucdavis. edu | *Ds* | 6 878 |
| 总计 | | | 140 097 |

IRRI 已经收集到几个不同遗传背景水稻的化学和射线诱导的大量突变体（表 27-3）。最大量的突变体来自 IR64 这个亚洲热带印度属水稻的最常见的变种。这个突变体库中，使用了多种诱变剂以产生不同大小片段的遗传损伤和突变，以适合正向和反向遗传学研究对突变体的需求（Wu et al. 2005）。至 2006 年 11 月，已经建立了超过 50 000个 $M_4$ 株系。IRRI 已经获得了充足的种子储备以作为 IRRI 种质资源库的永久资源进行保藏、传播应用。这个突变体库中遗传变异丰富，包括表型变异（～8%）和条件性状变异（如获得或失去对生物和非生物胁迫的抗性）。随着先进的基因组阵列技术的发展，化学和射线诱导的突变体的应用增长很快。缺失的序列可以通过常规的芯片技术被检测出来（Borevitz et al. 2003）。水稻中，Affymetrix 提供的覆盖全基因组的寡核苷酸芯片（每个基因有多个寡核苷酸序列探针），可以特异性用于探查水稻突变体的缺失序列区域（M. Bruce et al.，unpubl）。例如，之前所提到的，TILLING 可用来检测点突变引起的小片段的缺失/插入。

表 27-3 2006 年 2 月，IRRI 收集的辐射和化学诱变的突变体

| 基因型 | 诱变剂 | 基因损伤情况 | 主要用途 | 株系数目 |
|---|---|---|---|---|
| IR64 | 快中子(60 Gy) | 基因大量缺失、易位 | 正向遗传学阵列或 PCR 检测缺失 | 8 073 |
| | γ 射线（250、500 Gy） | 1～2 kb 的缺失、点突变 | 正向遗传学阵列或 PCR 检测缺失 | 15 295 |
| | 双环氧丁烷 | kb 范围的缺失、点突变 | 正向遗传学阵列或 PCR 检测缺失 | 16 520 |
| | EMS | 点突变 | 正向遗传学 TILLING | 12 539 |
| 总计 | | | | 52 427 |
| 其他基因型 | | | | |
| 章山黄(音)No. 2 | 1.6%EMS | | 抗稻瘟病 | 7383$M_3$ |
| Jinbubyeo | 0.6%EMS | | 耐寒 | 3000$M_3$ |
| FL478 | 1%EMS | | 耐盐性 | 11000$M_1$ |
| IRBB61 | 1%EMS | | 抗白叶枯病 | 11000$M_2$ |

## 活性重组构建基因型多样性

某基因座位的天然或诱导的等位基因变异提供了区域，在这个区域可以通过重组构建数量可观的基因型多样性。尽管等位基因多样性的确定需要很长时间，但是通过重组构建技术能够在几个世代之内获得数目庞大的基因型差异的个体用于多样性研究。能够进行最大化重组的遗传材料对于确定基因的功能是重要的资源。许多类型的分离群体在遗传作图和植物育种方面都是非常有用的，包括 RIL、染色体片段代换系、基因渗入系。从这些分离材料中，可以获得不同的特异遗传资源，如近同基因系和异质近交家族（如来自目的位点时异质性的近交系的姊妹品系）。

RIL 是遗传作图实验的支柱。稳定的优良品系可以永久保留，在重复的杂交定位实验中提供多种表型分析的选择。这特别是在检测微小的表型效应上有重要意义的。为进一步表型一致化和最大化地利用来自大量变种的 SNP 数据，我们已经开始了一半双等位基因的杂交计划，在所有重新测序的品种间进行两两配对杂交。由于育性的阻碍并不是所有的杂交都会成功，但是，那些来自于高密度基因型定位品系的 RIL 将对数量性状位点的分析提供特殊的信息。

## 转录图谱

因为水稻基因组测序的完成，使得以每一个染色体为单位描述基因表达模式绘制转录图谱成为可能。Kikuchi 等（2007）汇总了不同的可用于转录组分析的平台。22K 寡核苷酸阵列（Agilent）已经被用来检测激素应答和多种胁迫响应的转录调控研究（K. Satoh and Kikuchi，NIAS；R. Muthuraian and R. Mauleon，IRRI，未发表），可以期待能在一张芯片上扩展到 44K（Kikuchi et al.　2007）。北京基因组研究中心和耶鲁大学已经制作出 60K（双面）的寡核苷酸微阵列芯片，用于不同发育时期的全基因组表达谱的研究（Ma et al.　2005）。45K 的寡核苷酸微阵列芯片（第三版）由美国国家科学基金支持研发出来，加利福尼亚大学戴维斯分校负责共享使用的管理。50K 的 Affymetrix 芯片已经成为商业化的产品。与上述提到的长寡核苷酸链（60～70 个核苷酸残基）芯片不同，Affimetrix 芯片使用含 25 个残基的寡核苷酸链，设计检测每个基因的一组探针为 11 个寡核苷酸链。这些平台在寡聚核苷酸链的设计、基因组的覆盖率和花费上都是不同的。通过实验目的和计划费用，来选择所适合的平台。由于水稻微阵列的寡核苷酸芯片几乎覆盖了全部基因组范围，使得检测基因顺序相关的表达成为可能，其他方法只能在全部基因组和比邻序列完成后才有可能。

不同于依赖 DNA 的图谱与相应的表型一致而且是稳定的，转录图谱是动态的，取决于生物体发育阶段和检测时所处于的环境因素。这使得转录组分析很容易随着实验条件的变化而波动。转录图谱提供独特的“分子表型”，这对于理解在基因组水平上的遗传调控和搭建基因和表型之间的联系桥梁有很大贡献。耶鲁大学水稻图谱工程项目，已经构建出来自正常生长条件下印度属和日本属代表性水稻 40 种细胞类型的转录组图谱，

基本上代表了水稻不同器官和不同发育时期的状况（http://bioinformatics.med.yale.edu/rc/overview.jspx）。这个丰富的水稻数据库为进行来自其他研究的转录组数据提供了一个基本的参考。

已有证据表明基因表达模式受其在染色体上的定位影响。动植物转录组学分析的结果都显示出染色体区域和邻近基因表达关联的证据（Caron et al. 2000；Cohen et al. 2000；Spellman and Rubin 2002；Doss et al. 2005；Petkov et al. 2005；Williams and Bowles 2004；Jiao et al. 2005；Zhan et al. 2006）。在水稻中，Ma 等（2005）报道说基因组大约有 10%基因显示出关联性表达，这与在拟南芥中的结果是一致的（Zhan et al. 2006）。NIAC（国家农业科学研究所）和 IRRI 合作进行在胁迫应答反应中的特异转录图谱。建在 22K 寡聚核苷酸芯片基础上，病原体和水分胁迫下的水稻丰富基因表达数据库已完成。特别有趣的是，观测得到了基因表达模式和基因的染色体座位之间的关系。例如，在两个品种 Apo（耐旱型）和 IR64（干旱敏感型）之间，进行营养生长期的水分胁迫的比较分析，我们发现在干旱胁迫下 Apo 品种中的 14 个关联表达区域（20 基因窗口）(R. Muthurajan，K. Satoh，R. Mauleon，IRRI and NIAS，unpubl.)。一个区域位于 1 号染色体（38.5～39Mb）与一个报道过的耐旱 QTL 区共定位，第二个区域位于 8 号染色体的（22.5～23.5Mb）与另一个耐旱 QTL 区共定位。关联表达区域分析的应用使我们能够具体确定染色体受 QTL 影响区域（10～20 个基因），尽管差异表达不是那么明显。转录图谱显示差异表达基因和关联表达区域可被用来初步确定 QTL 获选区域的名单。我们评测一种可能性，即多种相邻基因的共表达可能属于同一个 QTL 区，就可以在实际中用期望的表型来筛选基因型。

作为来源寡核苷酸序列芯片的转录组学数据的补充，条件特异性表达数据可以用 SAGE（基因表达系列分析）和 MPSS（massively parallel signature sequencing）技术来进行（Gowda et al. 2006）。这些数据锚定在假定的分子上可以产生丰富的基因表达信息来源。蛋白质组学数据可以用类似方式分析（Komatsu and Yano 2006）以获得额外的信息，明确染色体结构在基因表达乃至表型上所起的作用。

## 实践筛选

水稻中用于基因图谱的分子标记已经有很长时间了，利用这些分子标记已经完成了具有重要经济价值的基因的遗传学图。但是，这些基因在分子标记辅助筛选（marker-assisted selection，MAS）的育种中没有得到进一步的应用。基因作图和 QTL 定位在 MAS 实践中仅仅是第一步，还需要完成许多步骤后才能真正用于育种。鉴定和使用分子标记用于筛选之前，还需要在不同的遗传群体中验证基因或 QTL 的效应，进一步精细遗传作图，以及建立与基因两翼序列紧密连锁的多样性分子标记的工具箱（Langridge et al. 2001；Collard and Mackill 2007）。

对于重要基因和作用显著的 QTL 很容易确定用于 MAS 筛选的分子标记。这些基因即使不借助于分子标记，在品种培育过程中也很容易操作。但是不管怎样，MAS 在单个植株的相同性状主效基因的金字塔中处于顶尖优势。控制的同一性状的基因经常会

有上位作用；因此，一个基因的表型可能会掩盖其他表型性状。这已被作为一种策略从高抗病基因型中获得多重抗性基因，如水稻白叶枯病（Huang et al. 1997；Joseph et al. 2004）和稻瘟病（Hittalmani et al. 2000；Tabien et al. 2000；Fjellstrom et al. 2004）。

对于主基因来说MAS的另一个用途是通过主基因的选择性转化或利用QTL的标记辅助的回交（MAB）来增加已有的品种变异。在植物分子标记技术发展的早期，人们仅仅认识到可以用标记进行感兴趣的小片段染色体的转化（Tanksley et al. 1989）。模拟研究显示遗传标记在品种改良中可促进回交速率和减少连锁的拖累（Frisch et al. 1999a，b；Hospital 2001；Servin and Hospital 2002）。基因两翼序列标记可用于重组子的筛选能明显使外来导入基因两侧的染色体片段减小（Frisch et al. 1999b；Hospital 2001）。Chen等（2000，2001）利用这种方法将包含白叶枯病抗性基因 *Xa21* 的小染色体片段转化到在中国广泛利用的两个杂交亲本。这种方法比起一般的杂交品系需要更大的回交群体，因此这种方法仍未被水稻育种者广泛的采用。不管怎样，最近在将涝灾耐受基因 *Sub1* 位点转化广泛种植的品种取得的成功，将会鼓舞人们更多地采用这种方法（Xu et al. 2006）。

QTL在MAS应用中依然存在一些困难，因为精细作图很难，并且作为单因素操作时效应较低。另外，由于存在着遗传背景因素效应和与环境的相互作用限制了QTL的应用。由于这些原因，高效的QTL对于MAS或MAB是更适合和需要的（Holland 2004；Mackill 2006）。QTL图谱通常会存在许多错误，图谱的精确性很大程度上取决于作图所使用群体数量的大小（Beavis 1998）。利用作图群体的后代或其他不同的遗传作图群体验证QTL图谱的研究是特别重要的。不管怎样，通过验证的QTL通常是非常精确的，就如同最近被克隆了的QTL所展示的那样（Price 2006）。

具有重大效应的QTL可用主基因同样的方法进行操作，虽然在水稻研究中的例子是有限的。Steele等（2006）在根的性状中提供了一个例子，但是应该注意到许多根中的QTL转化不同遗传背景的植株没有得到表达。对于微效的QTL，可能不适用于传统的MAS方法。获得好的精细QTL图谱是相当困难的，同时遗传背景的影响也是很高的。对于这些类性状，了解这些性状的自然特征和揭示QTL背后的基因将是非常有用的。微效的稻瘟病抗性基因研究显示水稻宿主防御基因经常位于QTL位点附近，这些基因的等位基因也于抗性水平相关（Liu et al. 2004；Wu et al. 2004）。结合遗传图谱和转录组分析可为揭示候选防御基因的作用提供进一步证据（R. Maulelon and B. Liu，IRRI unpubl.）。最后，突变分析和RNA干扰研究，在候选基因的表达被抑制的条件下，证明了一些QTL控制下的抗性基因的作用（P. Manosalva. unpubl.）。

在回交中遗传背景筛选的应用和多个基因同时操作的追求，导致对利用MAS进行全基因组有效筛查的需求。利用SSR标记，每个染色体臂至少要有4个标记可用于选择相应的供体染色体（Whittaker et al. 1996；Servin and Hospital. 2002），这仍可能遗漏带有双重重组子的小区域。带有自动测序装置的多元SSR标记是应用在遗传背景筛选的更有效的方法（Coburn et al. 2002）。但是，芯片类的方法可以检测到SNP或SFP（单一特征多态性）很可能更适合于对整个基因组的检测分析。以芯片为基础的整

个基因组的 SNP 和 SFP 的检测分析（Borevitz et al. 2003；Hazen and Kay 2003），将有可能替代用于 MAS 的 SSR 标记，因为后者在分离群体中的全基因组检测能力非常有限。

最近，D. Galbraith 和亚利桑那州立大学的合作者完成了寡聚核苷酸探测的设计，可用于水稻全基因组的 SFP 检测。SFP 检测微芯片的唯一特征是相对较低的消费成本，使得它在追溯育种材料遗传背景的常规技术的应用成为可能。进一步的降低成本，将会为建立储藏于 IRRI 的国际基因库的大于 102 000 份种质的遗传学“条形码”标签奠定商业化的可行性。随着转录分析成本的下降，如同在最近的拟南芥研究中展示的那样利用转录图谱在不同条件下选择基因型也会实现（Kliebenstein et al. 2006）。从作图结果的整合，表达分析和选择响应都将为有效地利用水稻等位基因和基因型多样性提供强有力的基础。

## 结论

可利用的全基因组序列为认识和利用遗传多样性开辟了新途径。在染色体背景的所有基因序列的揭示对于阐述基因组调控是十分必要的基础。水稻遗传多样性的分析将超越等位基因变异样本，包括在 DNA 和表达两个层次水平上的整合分析。这将能帮助我们理解 DNA 序列变异是如何通过全基因组的响应最终在表型上表现出差异。这个新的角度对于在农作物上理解遗传变异和遗传多样性是空前的。这可能在揭示许多农业性状的深层次控制机理，特别是那些目前为止依然是很难确定或在育种实践中很难操作的性状。

可以预计随着基因组测序、表达分析花费的减少，基因组变异和表达的多态性能被完全证实。在接下来几年里，限制因素将是用于设计高质量作图和基因身份鉴定分析的遗传资源的可使用性。我们如何才能利用全新高效技术去沉默或活化靶基因或基因簇？我们如何利用重组技术构建理想的重组基因或者能在整株植物或作物水平上展现特定功能的基因组呢？这些仅仅是新基因组技术和全基因组分析所希望解决的问题的一部分。

最后指出，只有当恰当的性状评价系统建立之后，应用基因组信息进行天然或诱导变异体的功能识别的潜力才能实现。许多农业上重要的性状是有条件的，如生物和非生物胁迫的抗性。在 IRRI 中，小的核心层变异品种的表型分析正在进行中，但是，大量收集工作对于在不同环境和地域进行表型的评价依然是必要的。INGER 提供了一种机制，但是一个主要挑战是努力构建一个能产生高质量和广泛可用数据的表型分析网络。另一个挑战是怎样对遗传资源进行灵活的改变使之能很好地进行独立检验。因此，共享数据和遗传资源的自由流通对于水稻遗传多样性及其应用的探索起着至关重要的作用。

## 致谢

我们感谢那些未署名的同事审阅书稿并提供未公开发表的信息。感谢 Ariel Javellana 为设计和制作图 27-2 所作的贡献。一些未公开的工作引用自美国农业部野村综合研

究所（USDA NRI）和时代挑战项目（Generation Challenge Program）资助的合作项目。

# 参考文献


Ammiraju J.S.S., Luo M., Goicoechea J.L., Wang W., Kudrna D., Mueller C., Talag J., Kim H.R., Sisneros N.B., Blackmon B., et al. 2006. The *Oryza* bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. *Genome Res.* **16:** 140–147.

Beavis W.D. 1998. QTL analysis: Power, precision and accuracy. In *Molecular dissection of complex traits* (ed. A.H. Paterson), pp. 145–162. CRC Press, Boca Raton, Florida.

Bhat R., Upadhyaya N.M., Chaudhury A., Raghavan C., Qiu F., Wang H., Wu J., McNally K., Leung H., Till B., et al. 2007. Chemical and irradiation induced mutants and TILLING. In *Rice functional genomics: Challenges, progress and prospects* (ed. N.M. Upadhyaya), pp. 151–186. Springer, New York.

Borevitz J.O., Liang D., Plouffe D., Chan H.-S., Zhu T., Weigel D., Berry C.C., Winzeler E., and Chory J. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13:** 513–523.

Brar D.S. and Khush G.S. 1997. Alien introgression in rice. *Plant Mol. Biol.* **35:** 35–47.

Caron H., van Schaik B., van der Mee M., Bass F., Riggins G., van Sluis P., Hermus M.-C., van Asperen R., Boon K., Vaûte P.A., et al. 2000. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Chen S., Lin X.H., Xu C.G., and Zhang Q.F. 2000. Improvement of bacterial blight resistance of 'Minghui 63', an elite restorer line of hybrid rice, by molecular marker-assisted selection. *Crop Sci.* **40:** 239–244.

Chen S., Xu C.G., Lin X.H., and Zhang Q. 2001. Improving bacterial blight resistance of '6078', an elite restorer line of hybrid rice, by molecular marker-assisted selection. *Plant Breed.* **120:** 133–137.

Cheng C., Motohashi R., Tsuchimoto S., Fukuta Y., Ohtsubo H., and Ohtsubo E. 2003. Polyphyletic origin of cultivated rice: Based on interspersion pattern of SINES. *Mol. Biol. Evol.* **20:** 67–75.

Coburn J.R., Temnykh S.V., Paul E.M., and McCouch S.R. 2002. Design and application of microsatellite marker panels for semiautomated genotyping of rice (*Oryza sativa* L.). *Crop Sci.* **42:** 2092–2099.

Cohen B.A., Mitra R.D., Hughes J.D., and Church G.M. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **26:** 183–186.

Collard B.C.Y. and Mackill D.J. 2007. Marker-assisted selection: An approach for precision plant breeding in the 21st century. *Philos. Trans. R. Soc. B Rev.* (in press).

Comai L., Young K., Till B.J., Reynolds S.H., Greene E.A., Codomo C.A., Enns L.C., Johnson J.E., Burtner C., Odden A.R., and Henikoff S. 2004. Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *Plant J.* **37:** 778–786.

Doss S., Schadt E.E., Drake T.A., and Lusis A.J. 2005. *Cis*-acting expression of quantitative trait loci in mice. *Genome Res.* **15:** 681–691.

Feltus F.A., Wan J., Schulze S.R., Estill J.C., Jiang N., and Paterson A.H. 2004. An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Res.* **14:** 1812–1819.

Feng Q., Zhang Y., Hao P., Wang S., Fu G., Huang Y., Li Y., Zhu J., Liu Y., Hu X., et al. 2002. Sequence and analysis of rice chromosome 4. *Nature* **420:** 316–320.

Fjellstrom R., Conaway-Bormans C.A., McClung A.M., Marchetti M.A., Shank A.R., and Park W.D. 2004. Development of DNA markers suitable for marker assisted selection of three Pi genes conferring resistance to multiple *Pyricularia grisea* pathotypes. *Crop Sci.* **44:** 1790–1798.

Frazer K.A, Wade C.M., Hinds D.A., Patil N., Cox D.R., and Daly M.J. 2004. Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 Mb of mouse genome. *Genome Res.* **14:** 1493–1500.

Frisch M., Bohn M., and Melchinger A.E. 1999a. Comparison of selection strategies for marker-assisted backcrossing of a gene. *Crop Sci.* **39:** 1295–1301.

———. 1999b. Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. *Crop Sci.* **39:** 967–975.

Garris A.J., McCouch S.R., and Kresovich S. 2003. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). *Genetics* **165:** 759–769.

Garris A.J., Tai T.H., Coburn J., Kresovich S., and McCouch S.R. 2005. Genetic structure and diversity in *Oryza sativa* L. *Genetics* **169:** 1631–1638.

Goff S.A., Ricke D., Lan T.H., Presting G., Wang R., Dunn M., Glazebrook J., Sessions A., Oeller P., Varma H., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296:** 92–100.

Gowda M., Jantasuriyarat C., Dean R.A., and Wang G.L. 2004. Robust-LongSAGE (RL-SAGE): A substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant Physiol.* **134:** 890–897.

Hazen S.P. and Kay S.A. 2003. Gene arrays are not just for measuring gene expression. *Trends Plant Sci.* **8:** 413–416.

Hinds D.A., Stuve L.L., Nilsen G.B., Halperin E., Eskin E., Ballinger D.G., Frazer K.A., and Cox D.R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.

Hirochika H., Guiderdoni E., An G., Hsing Y., Eun M.Y., Han C.D., Upadhyaya N., Ramachandran S., Zhang Q., Pereira A., et al. 2004. Rice mutant resources for gene discovery. *Plant Mol. Biol.* **54:** 325–334.

Hittalmani S., Parco A., Mew T.V., Zeigler R.S., and Huang N. 2000. Fine mapping and DNA marker-assisted pyramiding of the three major genes for blast resistance in rice. *Theor. Appl. Genet.* **100:** 1121–1128.

Holland J.B. 2004. Implementation of molecular markers for quantitative traits in breeding programs–challenges and opportunities. In *New directions for a diverse planet: Proceedings of the 4th International Crop Science Congress*, Brisbane, Australia. The Regional Institute Ltd., Gosford, Australia (www.cropscience.org.au).

Hospital F. 2001. Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. *Genetics* **158:** 1363–1379.

Huang N., Angeles E.R., Domingo J., Magpantay G., Singh S., Zhang G., Kumaravadivel N., Bennett J., and Khush G.S. 1997. Pyramiding of bacterial blight resistance genes in rice: Marker-assisted selection using RFLP and PCR. *Theor. Appl. Genet.* **95:** 313–320.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

International Rice Genome Sequencing Project (IRGSP). 2005. The map-based sequence of the rice genome. *Nature* **436:** 793–800.

Joseph M., Gopalakrishnan S., Sharma R.K., Singh V.P., Singh

A.K., Singh N.K., and Mohapatra T. 2004. Combining bacterial blight resistance and Basmati quality characteristics by phenotypic and molecular marker-assisted selection in rice. *Mol. Breed.* **13:** 377–387.

Jiao Y., Jia, P., Wang X., Su N., Yu S., Zhang D., Ma L., Feng Q., Jin Z., Li L., et al. 2005. A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. *Plant Cell* **17:** 1641–1657.

Khush G.S. 1997. Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.* **35:** 25–34.

Kikuchi S., Wang G.L., and Li L. 2007. Genome-wide RNA expression profiling in rice. In *Rice functional genomics: Challenges, progress and prospects* (ed. N.M. Upadhyaya), pp. 36–63. Springer, New York.

Kliebenstein D.J., West M.A.L., van Leeuwen H., Kim K., Doerge R.W., Michelmore R.W., and St. Claire D.A. 2006. Genomic survey of gene expression diversity in *Arabidopsis thaliana*. *Genetics* **172:** 1179–1189.

Komatsu S. and Yano H. 2006. Update and challenges on proteomics in rice. *Proteomics* **6:** 4057–4068.

Langridge P., Lagudah E.S., Holton T.A., Appels R., Sharp P.J., and Chalmers K.J. 2001. Trends in genetic and genome analyses in wheat: A review. *Aust. J. Agric. Res.* **52:** 1043–1077.

Leung H. and An G. 2004. Rice functional genomics: Large scale gene discovery and applications for crop improvement. *Adv. Agron.* **82:** 55–111.

Liu B., Zhang S.H., Zhu X.Y., Yang Q.Y., Wu S.Z., Mei M.T., Mauleon R., Leach J., Mew T., and Leung H. 2004. Candidate defense genes as predictors of quantitative blast resistance in rice. *Mol. Plant-Microbe Interact.* **17:** 1146–1152.

Lu B.R. 1999. Taxonomy of the genus *Oryza* (Poaceae): Historical perspective and current status. *Int. Rice Res. Notes* **24:** 4–8.

Ma L., Chen C., Liu X., Jiao Y., Su N., Li L., Wang X., Cao M., Sun N., Zhang X., et al. 2005. A microarray analysis of the rice transcriptome and its comparison to *Arabidopsis*. *Genome Res.* **15:** 1274–1283.

Mackill D.J. 1995. Classifying japonica rice cultivars with RAPD markers. *Crop Sci.* **35:** 889–894.

———. 2006. Breeding for resistance to abiotic stresses in rice: The value of quantitative trait loci. In *Plant breeding: The Arnel R. Hallauer International Symposium* (ed. K.R. Lamkey and M. Lee), pp. 201–212. Blackwell, Ames, Iowa.

Matsumoto T., Wing R.A., Han B., and Sasaki T. 2007. Rice genome sequence: The foundation for understanding the genetic systems. In *Rice functional genomics: Challenges, progress and prospects* (ed. N.M. Upadhyaya), pp. 5–21. Springer, New York.

McCallum C.M., Comai L., Greene E.A., and Henikoff S. 2000. Targeting induced local lesions IN genomes (TILLING) for plant functional genomics. *Plant Physiol.* **123:** 439–442.

McNally K.L., Bruskiewich R., Mackill D., Leach J.E., Buell C.R., and Leung H. 2006. Sequencing multiple and diverse rice varieties: Connecting whole-genome variation with phenotypes. *Plant Physiol.* **141:** 26–31.

Nakano M., Nobuta K., Vemaraju,K., Tej S.S., Skogen J.W., and Meyers B.C. 2006. Plant MPSS databases: Signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res.* **34:** D731–D735.

Olsen K.M., Caicedo A.L., Polato N., McClung A., McCouch S., and Puruggan M. 2006. Selection under domestication: Evidence for a sweep in the rice *Waxy* genomic region. *Genetics* **173:** 975–983.

Petkov P.M., Graber J.H., Churchill G.A., DiPetrillo K., King B.L. and Paigen K. 2005. Evidence of a large-scale functional organization of mammalian chromosomes. *PloS Genet.* **1:** 312–322.

Price A.H. 2006. Believe it or not, QTLs are accurate. *Trends Plant Sci.* **11:** 213–216.

Raghavan C., Naredo M.E.B., Wang H., Atienza G., Liu B., Qiu F., McNally K.L., and Leung H. 2007. Rapid method for detecting SNPs on agarose gels and its application in candidate gene mapping. *Mol. Breeding.* **19:** 87–101.

The Rice Chromosome 10 Sequencing Consortium. 2003. In-depth view of structure, activity and evolution of rice chromosome 10. *Science* **300:** 1566–1569.

The Rice Chromosome 3 Sequencing Consortium. 2005. Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverse grass species. *Genome Res.* **15:** 1284–1291.

Sasaki T., Matsumoto T., Yamamoto K., Sakata K., Baba T., Katayose Y., Wu J., Niimura Y., Cheng Z., Nagamura Y., et al. 2002. The genome sequence and structure of rice chromosome 1. *Nature* **420:** 312–316.

Second G. 1982. Origin of the genetic diversity of cultivated rice (*Oryza sativa* L.): Study of the polymorphism scored at 40 isozyme loci. *Jpn. J. Genet.* **57:** 25–57.

———. 1985. Evolutionary relationships in the *sativa* group of *Oryza* based on isozyme data. *Genet. Sel. Evol.* **17:** 89–114.

Servin B. and Hospital F. 2002. Optimal positioning of markers to control genetic background in marker-assisted backcrossing. *J. Hered.* **93:** 214–217.

Shen Y.J., Jiang H., Jin J.P., Zhang Z.B, Xi B., He Y.Y., Wang G., Wang C., Qian L., Li X., et al. 2004. Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol.* **135:** 1198–1205.

Spellman P.T. and Rubin G.M. 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1:** 5.

Steele K.A., Price A.H., Shashidhar H.E., and Witcombe J.R. 2006. Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. *Theor. Appl. Genet.* **112:** 208–221.

Sunkar R., Girke T., and Zhu J.K. 2005. Identification and characterization of endogenous small interfering RNAs from rice. *Nucleic Acids Res.* **14:** 4443–4454.

Tang T., Lu J., Huang J., He J., McCouch S.R., Shen Y., Kai Z., Purugganan M.D., Shi S., and Wu C.I. 2006. Genomic variation in rice: Genesis of highly polymorphic linkage blocks during domestication. *PLoS Genet.* **2:** 1824–1833.

Tabien R.E., Li Z., Paterson A.H., Marchetti M.A., Stansel J.W., and Pinson S.R.M. 2000. Mapping of four major rice blast resistance genes from 'Lemont' and 'Teqing' and evaluation of their combinatorial effect for field resistance. *Theor. Appl. Genet.* **101:** 1215–1225.

Tanksley S.D., Young N.D., Paterson A.H., and Bonierbale M.W. 1989. RFLP mapping in plant breeding: New tools for an old science. *BioTechnology* **7:** 257–264.

Vitte C., Ishii T., Lamy F., Brar D., and Panaud O. 2004. Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). *Mol. Gen. Genomics* **272:** 504–511.

Watanabe Y. 1997. Phylogeny and geographical distribution of genus *Oryza*. In *Science of the rice plant* (ed. T. Matsuo et al.), pp. 29–39. Food and Agriculture Policy Research Center, Tokyo, Japan.

Whittaker J.C., Thompson R., and Visscher P.M. 1996. On the mapping of QTL by regression of phenotype on marker-type. *Heredity* **77:** 23–32.

Williams E.J.B. and Bowles D.J. 2004. Coexpression of neighboring genes in *Arabidopsis thaliana*. *Genome Res.* **14:** 1060–1067.

Wing R.A., Kim H.-R., Goicoechea J.L., Yu Y., Kudrna D., Zuccolo A., Ammiraju J.S.S., Luo M., Nelson W., Ma J., et al. 2007. The *Oryza* Map Alignment Project (OMAP): A new resource for comparative genome studies within *Oryza*. In *Rice functional genomics: Challenges, progress and prospects* (ed. N.M. Upadhyaya), pp. 408–421. Springer, New York.

Wu J.L., Sinha P.K., Variar M., Zheng K.L., Leach J.E., Courtois B., and Leung H. 2004. Association between molecular markers and blast resistance in an advanced backcross population of rice. *Theor. Appl. Genet.* **108:** 1024–1032.

Wu J.L., Wu C., Lei C., Baraoidan M., Boredos A., Madamba R.S., Ramos-Pamplona M., Mauleon R., Portugal A., Ulat V., et al.

2005. Chemical- and irradiation-induced mutants of *indica* rice IR64 for forward and reverse genetics. *Plant Mol. Biol.* **59:** 85–97.

Xu K., Xia X., Fukao T., Canlas P., Maghirang-Rodriguez R., Heuer S., Ismail A.I., Bailey-Serres J., Ronald P.C., and Mackill D.J. 2006. *Sub1A* is an ethylene response factor-like gene that confers submergence tolerance to rice. *Nature* **442:** 705–708.

Yu J., Hu S., Wang J., Wong G.K., Li S., Liu B., Deng Y., Dai L., Zhou Y., Zhang X., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296:** 79–92.

Yu J., Wang J., Lin W., Li S., Li H., Zhou J., Ni P., Dong W., Hu S., Zeng C., et al. 2005. The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3:** e38.

Zhan S., Horrocks J., and Lukens L.N. 2006. Islands of co-expressed neighboring genes in *Arabidopsis thaliana* suggest high-order chromosome domains. *Plant J.* **45:** 347–357.

Zhu Q. and Ge S. 2005. Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. *New Phytol.* **167:** 249–265.

# 互联网资源

**http://bioinformatics.med.yale.edu/rc/overview.jspx** Yale Virtual Center for Cellular Expression Profiling of Rice. T. Nelson and X.-W. Deng, Yale Department of Molecular, Cellular & Developmental Biology (MCDB); H. Zhao, Department of Epidemiology & Public Health, Yale School of Medicine

**http://orygenesdb.cirad.fr** OryGenesDB, an interactive tool for rice reverse genetics. Centre de coopération internationale en recherche agronomique pour le développement

**http://rapdb.lab.nig.ac.jp/** Rice Annotation Project DataBase. Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, National Institute of Agrobiological Sciences, National Institute of Advanced Industrial Science and Technology, Japan Biological Informatics Consortium

**www.generationcp.org** Generation Challenge Programme, Science for Better Crops

**www.irgcis.irri.org:81/grc/irgcishome.html**

**www.iris.irri.org** International Rice Information System (IRIS) is the rice implementation of the International Crop Information System (ICIS) which is a database system that provides integrated management of global information on genetic resources and crop cultivars. This includes germplasm pedigrees, field evaluations, structural and functional genomic data (including links to external plant databases) and environmental (GIS) data.

**www.tigr.org/tdb/e2k1/osa1/** TIGR (The Institute for Genomic Research) Rice Genome Annotation Database

**www.omap.org** The Oryza Map Alignment Project. Principal investigators R.A. Wing, University of Arizona, Arizona Genomics Institute; S.A. Jackson, Purdue University; L.D. Stein, Cold Spring Harbor Laboratory; C. Soderlund, Unviersity of Arizona, Arizona Genomics Computation Laboratory

# 28 小　　鼠

Claire M. Wade and Mark J. Daly

*Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114; Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142*



## 简介

近交的哺乳动物是一类对于我们了解遗传学及基因组学有帮助的生物体。尤其是近交的小鼠，因为其价格便宜，相对来说容易饲养和喂食，而且能够适应人类的触摸，早已作为常用的哺乳动物研究材料。另外，有关基因组操作的培育策略以及方法是有现成描述的。美国缅因州 Bar Harbor 的 Jackson 实验室给研究者提供了超过 2700 种的小鼠，这也是小鼠作为实验材料普及的一个证据。2002 年，随着小鼠基因组序列分析工作的完成，一系列小鼠基因组资源的阵列被开发出来，成为协助小鼠进行研究的工具。虽然定向杂交和数量性状基因座（QTL）图谱是小鼠遗传事件中的标准工具，但是最近却显示杂交试验小鼠的遗传特性，可能会在未来的基因作图中因使用 SNP 单体型进行基因定位而出现很大转机。本章，我们将讨论小鼠研究历史中给予人们的启示和与小鼠基因图谱分析新方法有关的小鼠基因组资源。

### 小鼠历史对于杂交实验小鼠基因组的影响

普通的家鼠（*Mus musculus*）来自今天的印度及东南亚地区（Silver 1995），在现代通过人类迁移和农业实践而有机会传播。如果不是全部的话，绝大多数家鼠属于 *Mus musculus* 物种，这个种在世界各地不同的地区分化成了很多亚种。从科学的角度来看，尽管其他（如 *Mus musculus bactrianus*）亚种已知也是存在的，但是最重要的亚种是通过欧洲东部及亚洲扩散的 *Mus musculus musculus*；而 *Mus musculus domesticus* 则是通过欧洲西部传播的，是欧洲西部、美洲人和澳大利亚人殖民化的结果；*Mus musculus*

*castaneus* 来自东南亚；*Mus musculus molossinus* 来自日本。

因杂交而起源的近交系小鼠，其发展如同宠物小鼠以及观赏小鼠一样（更多指的是“奇特的”小鼠，因为它们有独特的皮毛和行为）。因此，在欧洲和亚洲，人们由于消遣而喂养小鼠，使得大量的小鼠被保留了下来。喜欢喂养和保存小鼠的人们努力创造了新的并且不同的表型。因为人们开始到世界各地去旅行，小鼠们也被交换了，这就足以使各地的 *Mus musculus* 一同产生出不同的亚种。

通常用于科学研究的近交小鼠品系的来源也是多样的。很多所谓的标准近交小鼠品系起源于 Clarence Cook Little 和 William Castle 的实验室，这个实验室的小鼠品系大多来自于同一个特别的小鼠的喂养者 Abbie Lathrop（Silver 1995）。几乎所有实验室的小鼠都是杂交产生的。这就可以为一个杂交品系提供额外的遗传稳定性，并省略了亲本染色体相的任何需要，因为所有的位点都是纯合的。最重要的是，完全纯合的基因组提供了可更新的动物研究策略，使得实验在跨越空间和时间的情况下可以被重复。另外一些主要的杂交系包括瑞士来源的小鼠系以及野生来源的杂交系，并不是来源于 Castle 和 Little 的品系。野生型来源的杂交鼠系可以代表不同的 *Mus musculus* 亚种（尤其是 *domesticus*、*musculus* 和 *molossinus*），或者说它们是另一个与 *Mus* 种（如 *Mus spretus*）有极大关系的品系。比起瑞士小鼠和 Castle-Little 小鼠，这些小鼠拥有更高水平的趋异性（Beck et al. 2000）。一般来说，它们更容易被认为是代表远系杂交种群中挑选出的各条染色体的单一拷贝。

## 基因组变异的观察

在成对小鼠品系的比较中，多态性不是随机地分布在基因组中的。正确的说法是，SNP 成簇地出现于高差异性区域或者低差异性区域。这已经被 Lindblad-Toh（2002）等证明了，他们观察到出现在表达序列标签（EST）中的 SNP 不能遵循所希望的泊松分布。无论 EST 中 SNP 的数量是比预先通过碱基检查给出的数量少还是多，差距都是显著的。

SNP 是国际小鼠基因组序列分析协会（International Mouse Genome Sequencing Consortium）小鼠基因组测序项目的一部分，它是通过少数杂交实验鼠（129S1/SvImJ、BALBc/ByJ、C3H/HeJ）之间的全基因组鸟枪测序比较（50 000 reads），以及和 C57BL6/J 品系的全序列比较发现的。通过分析确定的 SNP 基因组的形式（Wade et al. 2002），Lindblad-Toh 等（2000）的发现被证实。同时，Celera Discovery Corporation 通过使用 4 种小鼠品系（129X1/SvJ、DBA/2J、A/J 和 C57BL6/J）的小覆盖度（1.5×）以及 129S1/SvImJ 品系的更小覆盖度对小鼠进行了测序（Mural et al. 2002）。Richard Mural（pers. comm.）注意到品系中 SNP 的等效的、不均匀的分布。这个问题早已被 Wade 等（2002）甚至更早的 Bonhomme 等（1987）提出，他们认为 SNP 的这种不均匀分布取决于小鼠品系的杂交历史。在一对品种中观察到的低差异性区域会表明此区域中有共同的祖先历史，然而高差异性区域将会通过所研究的基因座位指出全然不同的亚种祖先。小鼠品系中主要的祖先来源是 *Mus musculus domesticus*，常见的次级祖先型是 *Mus musculus musculus*。

小鼠的拷贝数目多态性和插入/删除事件还没有被广泛研究。已经发表的最广泛的分析结果包括了这些使用比较基因组杂交技术来研究品系之间和插入/删除事件之间的拷贝数量的变化（Chung et al. 2004；Adams et al. 2005；Snijders et al. 2005）。一些研究已经检测到品系之间在复制数上的差异性。具体地，7 号染色体和 14 号染色体上的区域被列出举证。与活体小鼠相比，这些研究分析了同样的细胞株，这些结果很可能是在细胞株中传代的事件，而非在那个种的所有成员中可以发现的。小范围的插入、删除事件已经研究了 26kb 了（Ideraabdullah et al. 2004），结论是插入/删除事件的倾向很小（40%是单碱基事件），而它们似乎起源于亚种之间，而不是亚种内部。

## 小鼠比较基因组结构的完整性在性状作图实践中的影响

在数量性状位点（即性状是连续的，而非“全或无”的表型，如肥胖；对于普通的健康问题表现为易患病的体质，如心脏病和糖尿病）上有显著影响的基因一贯是难于鉴定的。过去，这些性状点的图谱已经被做出来，方法是通过鉴定对相应表型有不同易患性的遗传病个体的谱系或家族，接着通过交配产生的 $F_2$ 代后代来分离含有影响性状的因子的基因片段，筛选、分离是通过低密度的微卫星和 SNP 标记技术完成的。虽然这些常可以被认为是成功的表型相关鉴别标记，但是由于少数繁殖后代的低水平重组，要想将关联之间的距离缩短到数个 Mb 之内，是一件很困难的事情。

小鼠中的 SNP 变异方式预示着使用这些多态现象绘制 QTL 遗传图谱的可能性。到目前为止，小鼠的基因组还剩余一些相对未确定的区域，而这些区域的存在直接与品系的表型以及品系间的 SNP 变异相关联（首先由 Grupe 等于 2001 年提出）。计算机模拟作图（in silico mapping）的限制取决于大量有效相关的潜在统计学影响（Chesler et al. 2001）以及可用于比较的有限的已确定表型的品系数量（Pletcher et al. 2004）。这样的分析可能使得 QTL 的分辨率发生问题，并且限制了对病例的检测能力，这就难以实现通过遗传变异解释所有表型变异中有意义部分的目标。一些研究者还没有想要接受分析得很深入的研究方法，这些方法暗示了共同祖先的区域排他性（Yalcin et al. 2004），同样，偶然的序列差异可能会在其他一致性品系中被发现。这种方法已经被 Wade 和 Daly（2005）以及 Cervino 等（2006）提出，因为压倒多数的变异将会在基因组的间隔序列中容易识别的、仅仅覆盖基因组 50%的片段中被发现。研究者使用计算机模拟作图方法，在孟德尔位点作图中得到了很好的结果。然而，几乎可以肯定的是，复杂表型中的少数等位基因足以为有限的品系（如 50 种 Haplotype Map 品系的小鼠）解释足够多的表型变异，这样也就在全基因组相关检测例子中鉴定出了因果关系。这个问题因为如下原因被放大了：尽管允许相邻标记的从属关系以及品系之间自身的相互关联存在，但是有效检测已知表型和 SNP 单体型之间关联强度的统计学方法很少。由于许多近交的实验室品系来源相同，所以不可能想假设所有品系在基因组范畴内的遗传基础上都同样地没有关联（图 28-1）。在许多情况下，通过测验仅存在于已经知道基因区域 QTL 的某个基因组区域内的多态性，这些问题就会有实质性的减少。这就使得分析检测的包袱减少了，因此增加了研究检测有意义关联的能力。通过拥有相对长的品系间演化分支长度的表型分析品系，只分析已知拥有同一祖先的一组中的一个品系（如

C57/C58 组中的品系），品系之间的相互关联问题就能在很大程度上被解决。这一分析范例最新的成功实验结果会在下文介绍。

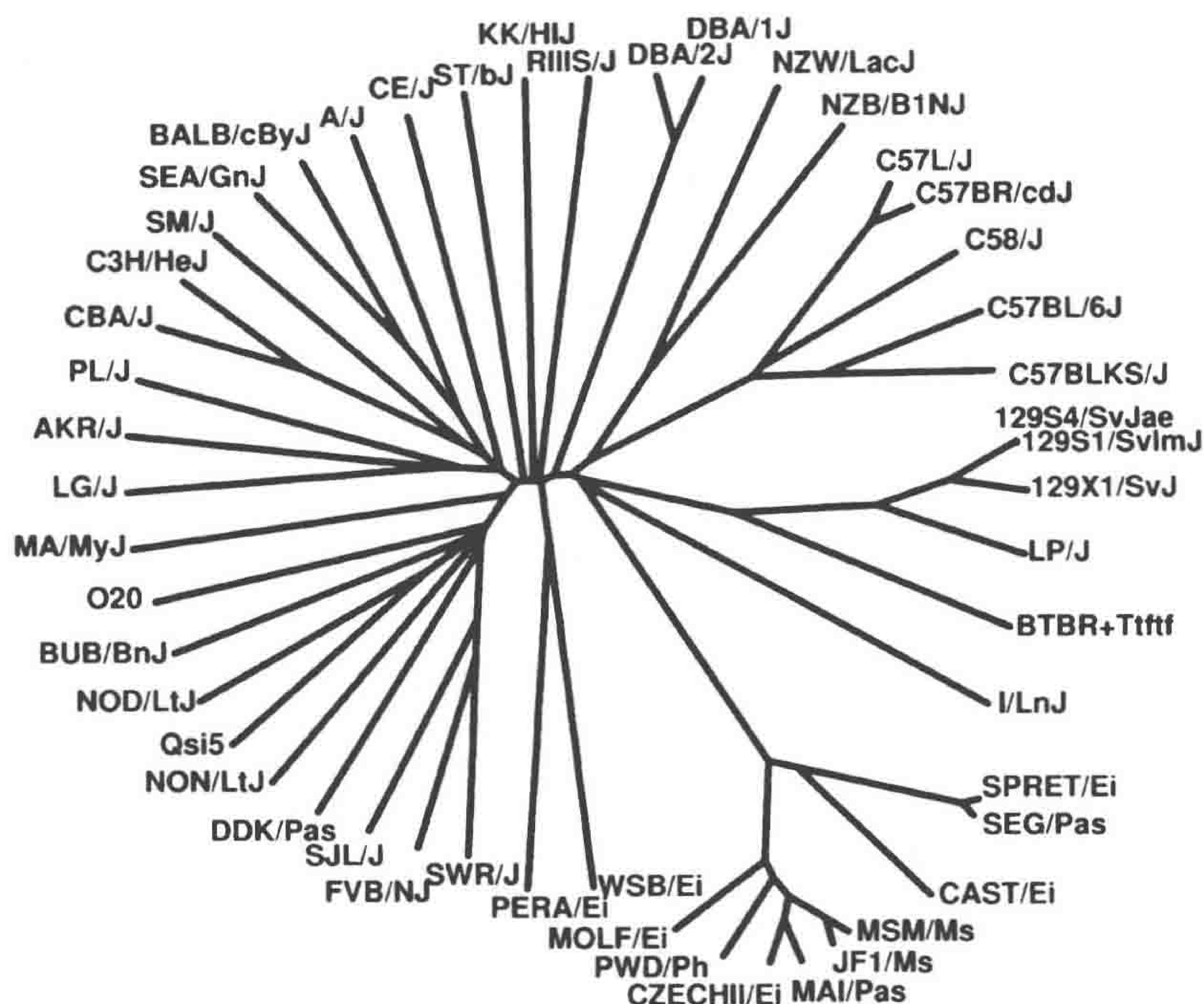



图 28-1　小鼠单体型作图计划中 49 个小鼠品系的基因组范围系谱分析，显示了密切联系的品系的聚类结果。相对于野生来源的品系，纯系内的分支夸大了些，因为 SNP 被确定是用于研究典型品系间的多态性的

## 可供选择的基因作图资源

对于遗传学家来说，用实验小鼠进行实验是幸运的，有许多资源可以帮助探索复杂的表型中往往难以估量的遗传成分的贡献。国际小鼠基因组测序协会绘制的小鼠基因组序列草图在 2002 年 2 月发表，而相应的文章在同年 12 月刊登（Waterston et al. 2002)。早些时候，Celera 发现体系（Celera discovery system）发表了他们对小鼠 16 号染色体的独立分析结果（Mural et al. 2002)。自那时以来，小鼠基因组就已进入基因组修正的最后阶段。这意味着，所有已知的缺口在修正中都通过使用各种昂贵的但准确的技术进行了填补。伴随小鼠基因组分析过程而整理的文章，目前正在准备中。

在小鼠基因组测序项目中，近 400 000 个 SNP 向 dbSNP 进行了投稿（http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD = search&DB = SNP&term = txid10090%5Borgn%5D)。此后不久，通过一些其他团体，SNP 又增加了内容，包括（但不限于）Celera 发现体系、Roche 生物科学学会和 Sanger 协会 。最近，SNP 的数量在公共领域已经大大增加了，多达 650 万，这主要是由于最近通过 Perlegen 科学学会与美国国立环境与健康科学研究院（National Institute of Environmental and Health

Science）建立研究关系加紧了测序的开展（K. Frazer，pers. comm.）。这一序列分析的结果涉及15个近交系小鼠品系的微阵列。在阵列上包括的序列展示了小鼠基因组中的非重复性序列部分。所涉及的品系包括129S1/SvImJ、A/J、AKR/J、BALBc/ByJ、BTBR T＋tf/J、C3H/HeJ、DBA/2J FBV/NJ、KK/HIJ、NZW/LacJ、NOD/LtJ、CAST/EiJ、MOLF/EiJ、PWD/PhJ和WSB/EiJ。这些品系（野生衍生自交系）中的前4个不仅用于实验品系，而且被作为假定祖先系代表用来研究小鼠的祖先。

在公共领域大量SNP的可用性促进了国际上重大基因型成果的产生。基因型分析成果的一个焦点是描述常用于繁殖实验的小鼠品系的基因组特点。诺华研究基金会（Novertis Research Foundation）的基因组学研究所向小鼠研究协会中的合作者发布了在48个品系中检测的10 990个多态性位点的基因型（Pletcher et al. 2004）。确定基因型的品系是优先考虑的小鼠品系，这是由Jackson实验室通过参考小鼠研究协会的小鼠非增殖部数据库（Mouse Phenome Database）来界定的（Bogue 2003；Bogue and Grubb 2004）。2005年，Wellcome Trust发布了大量的基因型（见下文），其中包括许多近交实验室品系。2006年期间，哈佛大学和麻省理工学院的Broad研究所公布了从经典近交系（而不是野生来源的品系）分离的49个小鼠近交系中超过148 000个多态性部位的基因型（www.broad.mit.edu/personal/mjdaly/mousehapmap.html）。Perlegen Sciences的再测序结果导致了在其代表品系群体中潜在多态性部位的确定，并同时为实验测试的一套较小的品系提供基因型。用于小鼠基因单体型作图的已确定基因型的品系在表28-1中总结。

**表28-1　单体型作图分析中包括的小鼠品系**

（www.broad.mit.edu/personal/mjdaly/mousehapmap.html）

| 近交实验品系 | | 野生来源品系 |
|---|---|---|
| 129S1/SvImJ | I/LnJ | *Mus m. castaneus* |
| 129X1/SvJ | KK/HIJ | CAST/Ei |
| A/J | LG/J | |
| AKR/J | LP/J | *Mus m. musculus* |
| BALB/cByJ | MA/MyJ | CZECHII/Ei |
| BTBR＋Ttftf | NOD/LtJ | PWD/Ph |
| BUB/BnJ | NON/LtJ | |
| C3H/HeJ | NZB/B1NJ | *Mus m. molossinus* |
| C57BL/6J | NZW/LacJ | JF1/Ms |
| C57BLKS/J | O20 | MAI/Pas |
| C57BR/cdJ | PL/J | MOLF/Ei |
| C57L/J | Qsi5 | MSM/Ms |
| C58/J | RIIIS/J | |
| CBA/J | SEA/GnJ | *Mus m. domesticus* |
| CE/J | SJL/J | PERA/Ei |
| DBA/1J | SM/J | WSB/Ei |
| DBA/2J | ST/bJ | |
| DDK/Pas | SWR/J | *Mus spretus* |
| FVB/NJ | | SEG/Pas |
| | | SPRET/Ei |

对于大量的重组近交系小鼠品系，基因型已通过 Wellcome-CTC Mouse Strain SNP Genotyped 建立的网站公开发布（http://www.well.ox.ac.uk/mouse/INBREDS）。已测定 13 377 个 SNP 的 480 个品系包括重组近交系、2300 个杂合库（Herterogeneous Stock）小鼠、许多实验室近交系和野生衍生近交系。杂合库来自多对 $F_2$ 贡献系杂交后代的后代。

在某种程度上，为了克服当前近交系在数量和遗传多样性方面的限制，正在对一些可以协同加强这些遗传资源的重要新鼠资源进行培育。小鼠染色体代替系最近已被发展为 A/J 品系和在 C57BL6/J 背景下的 129S1/SvImJ。A/J 染色体代替系随时可用（Nadeau et al. 2000；Singer et al. 2004），但 129 型品系仍处于发展中。这些品系通常包括一个 C57BL6 /J 遗传背景，它带有一个来自代替品系的单染色体。这种品系能够在一个染色体接着一个染色体的基础上解剖复杂的性状（Petryshen et al. 2005）。其发展要求许多代杂交小鼠的育种和基因型分析。

复杂性状协会（Complex Trait Consortium）是一个组织，旨在促进可用于理解人类疾病的资源的开发。作为这项倡议的一部分，该小组正在在小鼠中发展“交叉合作”（the collaborative cross，Churchill et al. 2004）。这是为了为哺乳动物复杂疾病的综合分析提供可作为共同参照的一组不同个体。这个协会的目标是生产出大量能普遍使用的、经过遗传学设计的一系列小鼠。一旦小鼠被进行一次基因型分析，单个品系都将会保留而库存起来；小鼠遗传学研究者购买了它们后，可以对其进行表型分析。

这些资源最终的关键部分是 Jackson 实验室（Bar Harbor，Maine）正在进行的小鼠表型数据库研究项目（Bogue 2003；Bogue and Grubb 2004）。这个项目包括了实验用杂交小鼠的大量标准化表型分析。表型分析实验的结果可以与小鼠协作组内其他成员共享。可行的数据不仅能够被普遍用于近交系中，还能被用于很多重组的近交品系间。表 28-2 总结了在网上发表的小鼠性状遗传定位资源。

**表 28-2　网上可获得的小鼠资源**

| 资源类型 | 包含的小鼠品系 | URL | 主要研究者 |
|---|---|---|---|
| 品系基因型——148 000 基因座位加性分析 | 50 个近交实验品系 | http://www.broad.mit.edu/personal/mjdaly/mousehapmap.html | Mark Daly |
| 品系基因型——～10 990 基因座位;也用作图方法 | 重组近郊系加上近交实验品系 | http://www.well.ox.ac.uk/mouse/ | Richard Mott |
| 品系基因型——高密度 | 15 个近交实验品系 | http://mouse.perlegen.com | Kelly Frazer |
| 品系和作图资源 | 大多数商用品系 | http://www.jax.org | Ken Paigen |
| SNP 数据库 | 迄今为止的所有已经确定基因型的 | http://www.ncbi.nlm.nih.gov/SNP/ | National Institutes of Health |
| 小鼠基因组资源(Mouse genome resource) | C57BL/6J | http://www.ncbi.nlm.nih.gov/genome/guide/mouse/ | National Institutes of Health |
| 表型数据库 | 50 个常见品系 | http://www.jax.org/phenome | Molly Bogue |
| 基因组浏览 | C57BL/6J | http://genome.ucsc.edu | Jim Kent |

续表

| 资源类型 | 包含的小鼠品系 | URL | 主要研究者 |
|---|---|---|---|
| 基因组浏览 | C57BL/6J | http://www.ensembl.org | Ewan Birney |
| 基因组浏览和作图工具 | 所有品系 | http://www.informatics.jax.org | Jackson Laboratory |
| 复杂性状作图资源 | 所有实验品系 | http://www.complextrait.org | Complex Trait Consortium |
| 通过血统作图进行遗传鉴定 | 已经确定基因型的品系 | http://mouseibd.florida.scripps.edu | Allesandra Cervino |
| 关联作图 | 表型组(phenome)品系 | http://snp.ucsd.edu/mouse | Eleazar Eskin |

## 着手计划小鼠复杂性状定位实验

研究者采取的性状作图分析方法依赖于性状所包含的信息量，以及对于性状的作图分析来说其遗传结构的预期复杂度。

### 实例一：已知的 QTL

先前通过连锁检测到 QTL 的地方，利用一种电子杂交定位技术可以缩短关联之间的间距。要做这个工作，研究者应该了解 QIL 品系所有的有效基因型。如果同样的 QTL 已经在多重杂交中被观察到，那么上述过程会大大地加快。在间隔中，研究者将识别预期含不同性状的品系，因为此性状具有不同的祖先单体型（如在遗传上异质的）的区域。相对来说没有争议的假设是，可以对那些已知使得定位品系在遗传学上产生差异的基因作更进一步的评价。对于一对品系，仅这个独自的过程就能使间隔缩短高达 50%～60%（Wade et al. 2002)。在最初连锁区域中基因型的密度会被最初用作 QTL 定位的品系严重影响。Perlegen 再测序数据表现的品系会有最大的覆盖密度，随后产生 48 种高优先度的品系。

一旦通过这一过程已经确定了一系列可能的、最窄的间隔，研究者就可以对高保守的哺乳动物序列中外显子以及片段中无外显子的位点进行重新测序。另一个办法，研究者可以进行表达分析（Mehrabian et al. 2005；Drake et al. 2006）从而识别在间隔中的基因，因为在作图品系中存在着不同的表达。

如果通过这些方法在大一些的间隔中没有发现基因的话，定位的性状就可能会是共同遗传学背景下最新的突变，或者是由插入/删除多态性所引起的。最近的研究（McCarrol et al. 2006）已经确定插入/删除多态性通常会因祖先单体型的高连锁不平衡而分离。因此，这种类型的多态性很容易存在于出现 SNP 变异的间隔中，而 SNP 的变异又预示着很多的差异。当单体型相似度存在于关联间隔上，就说明对存在致密的基因分型数据的其他品系进行基因型分析是有用的，而高密度数据能够识别关联区域的终点。跨越这些品系来看，连锁的不平衡长度大约为 100 kb（Frazer et al. 2004)，比品系对之间的长度（接近 1Mb）要相对短一些。最近的一些研究表明伴随着 QTL 定位的电子

关联技术可能是一个有效的位置克隆技术。Liu 等（2006）通过在 21 个品系间一个编码 SNP 的关联确定了一个新的肺癌易感性基因。在剧烈的阿霉素诱导的肾病品系中，Zheng 等（2006）确定了一个 1.3Mb 的独特单体型，其中插入了一个明显连锁的 QTL。这就大大减轻了接下来获得连锁结果的负担。

实例二：非预存的信息

如果没有性状的已知显著连锁的区域，首先就是确定多个存在高密度基因型数据的代表品系的表型。如果可能的话，应该避免野生型来源的近交品系，因为它们不能提供有限变异的相同嵌合模式，而这些变异类型是来自 Castle-Little 实验室小鼠和 Swiss 小鼠的近交实验品系所共享的。在标准的近交品系中，只包括一个关系较近的家族中（如 C57、C58 小鼠家族）的一个代表是合适的，因为这些小鼠通常在遗传和表型是相似的。一旦表型检查鉴定出在性状上有分歧表型的小鼠，就很有可能使用一些技术为将来的研究确定间隔，如候选基因组合分析、电子杂交定位分析和表达分析。

# 结论

尽管已经测序的哺乳动物的数量在不断地增加，但是都很少能像小鼠那样在遗传研究中具有如此强大的优势。来自研究小鼠的实验室的大量可利用资源提供了一个探索哺乳动物功能性变异的无法超越的机会。这些方法更利于开发存在的遗传学定位结果，也可以作为共同的参考资源，定位影响复杂性状的新间隔。鉴定显著关联区域的最有效方法是多种技术的联合，包括单体型分析和表达分析技术。

# 参考文献

Adams D.J., Dermitzakis E.T., Cox T., Smith J., Davies R., Banerjee R., Bonfield J., Mullikin J.C., Chung Y.J., Rogers J., and Bradley A. 2005. Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. *Nat. Genet.* **37:** 532–536.

Beck J.A., Lloyd S., Hafezparast M., Lennon-Pierce M., Eppig J.T., Festing M.F., and Fisher E.M. 2000. Genealogies of mouse inbred strains. *Nat. Genet.* **24:** 23–25.

Bogue M. 2003. Mouse Phenome Project: Understanding human biology through mouse genetics and genomics. *J. Appl. Physiol.* **95:** 1335–1337.

Bogue M.A. and Grubb S.C. 2004. The Mouse Phenome Project. *Genetica* **122:** 71–74.

Bonhomme F., Guenet J.-L., Dod B., Moriwaki K., and Bulfield G. 1987. The polyphyletic origin of laboratory inbred mice and their rate of evolution. *J. Linn. Soc.* **30:** 51–58.

Cervino A.C., Gosink M., Fallahi M., Pascal B., Mader C., and Tsinoremas N.F. 2006. A comprehensive mouse IBD database for the efficient localization of quantitative trait loci. *Mamm. Genome* **17:** 565–574.

Chesler E.J., Rodriguez-Zas S.L., and Mogil J.S. 2001. In silico mapping of mouse quantitative trait loci. *Science* **294:** 2423.

Chung Y.J., Jonkers J., Kitson H., Fiegler H., Humphray S., Scott C., Hunt S., Yu Y., Nishijima I., Velds A., et al. 2004. A whole-genome mouse BAC microarray with 1-Mb resolution for analysis of DNA copy number changes by array comparative genomic hybridization. *Genome Res.* **14:** 188–196.

Churchill G.A., Airey D.C., Allayee H., Angel J.M., Attie A.D., Beatty J., Beavis W.D., Belknap J.K., Bennett B., Berrettini W., et al. 2004. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.* **36:** 1133–1137.

Drake T.A., Schadt E.E., and Lusis A.J. 2006. Integrating genetic and gene expression data: Application to cardiovascular and metabolic traits in mice. *Mamm. Genome* **17:** 466–479.

Frazer K.A., Wade C.M., Hinds D.A., Patil N., Cox D.R., and Daly M.J. 2004. Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 Mb of mouse genome. *Genome Res.* **14:** 1493–1500.

Grupe A., Germer S., Usuka J., Aud D., Belknap J.K., Klein R.F., Ahluwalia M.K., Higuchi R., and Peltz G. 2001. In silico mapping of complex disease-related traits in mice. *Science* **292:** 1915–1918.

Ideraabdullah F.Y., de la Casa-Esperon E., Bell T.A., Detwiler D.A., Magnuson T., Sapienza C., and de Villena F.P. 2004. Genetic and haplotype diversity among wild-derived mouse inbred strains. *Genome Res.* **14:** 1880–1887.

Lindblad-Toh K., Winchester E., Daly M.J., Wang D.G., Hirschhorn J.N., Laviolette J.P., Ardlie K., Reich D.E., Robinson E., Sklar P., et al. 2000. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.* **24:** 381–386.

Liu P., Wang Y., Vikis H., Maciag A., Wang D., Lu Y., Liu Y., and

You M. 2006. Candidate lung tumor susceptibility genes identified through whole-genome association analyses in inbred mice. *Nat. Genet.* **38:** 888–895.

McCarroll S.A., Hadnott T.N., Perry G.H., Sabeti P.C., Zody M.C., Barrett J.C., Dallaire S., Gabriel S.B., Lee C., Daly M.J., and Altshuler D.M. (International HapMap Consortium). 2006. Common deletion polymorphisms in the human genome. *Nat. Genet.* **38:** 86–92.

Mehrabian M., Allayee H., Stockton J., Lum P.Y., Drake T.A., Castellani L.W., Suh M., Armour C., Edwards S., Lamb J., et al. 2005. Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat. Genet.* **37:** 1224–1233.

Mural R.J., Adams M.D., Myers E.W., Smith H.O., Miklos G.L., Wides R., Halpern A,. Li P.W., Sutton G.G., Nadeau J., et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661–1671.

Nadeau J.H., Singer J.B., Matin A., and Lander E.S. 2000. Analysing complex genetic traits with chromosome substitution strains. *Nat. Genet.* **24:** 221–225.

Petryshen T.L., Kirby A., Hammer R.P., Jr., Purcell S., O'Leary S.B., Singer J.B., Hill A.E., Nadeau J.H., Daly M.J., and Sklar P. 2005. Two quantitative trait loci for prepulse inhibition of startle identified on mouse chromosome 16 using chromosome substitution strains. *Genetics* **171:** 1895–1904.

Pletcher M.T., McClurg P., Batalov S., Su A.I., Barnes S.W., Lagler E., Korstanje R., Wang X., Nusskern D., Bogue M.A., et al. 2004. Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol.* **2:** e393.

Silver L.M. 1995. *Mouse genetics: Concepts and applications.* Oxford University Press, New York.

Singer J.B., Hill A.E., Burrage L.C., Olszens K.R., Song J., Justice M., O'Brien W.E., Conti D.V., Witte J.S., Lander E.S., and Nadeau J.H. 2004. Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* **304:** 445–448.

Snijders A.M., Nowak N.J., Huey B., Fridlyand J., Law S., Conroy J., Tokuyasu T., Demir K., Chiu R., Mao J.H., et al. 2005. Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res.* **15:** 302–311.

Wade C.M. and Daly M.J. 2005. Genetic variation in laboratory mice. *Nat. Genet.* **37:** 1175–1180.

Wade C.M., Kulbokas E.J., III, Kirby A.W., Zody M.C., Mullikin J.C., Lander E.S., Lindblad-Toh K., and Daly M.J. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature* **420:** 574–578.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R,. Ainscough R., Alexandersson M,. An P., et al. (Mouse Genome Sequencing Consortium) 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Yalcin B., Fullerton J., Miller S., Keays D.A., Brady S., Bhomra A., Jefferson A., Volpi E., Copley R.R., Flint J., and Mott R. 2004. Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc. Natl. Acad. Sci.* **101:** 9734–9739.

Zheng Z., Pavlidis P., Chua S., D'Agati V.D., and Gharavi A.G. 2006. An ancestral haplotype defines susceptibility to doxorubicin nephropathy in the laboratory mouse. *J. Am. Soc. Nephrol.* **17:** 1796–800.

# 29 大　鼠

Edwin Cuppen,[1] Norbert Hübner,[2] Howard J. Jacob,[3] and Anne E. Kwitek[3]

*[1]Hubrecht Laboratory, Utrecht, The Netherlands; [2]Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Germany; [3]Medical College of Wisconsin, Milwaukee, Wisconsin 53226*



## 简介

在生理学、毒理学以及神经生物学研究领域中，实验用大鼠（*Rattus norvegicus*）是一种广泛使用的研究模式生物。同样它也是药物研究的最初阶段使用的一种主要动物模型。通过实验大鼠的生物学特征，能够鉴定多于 500 个的品系（http://rgd.mcw.edu/strains/），它们中的大部分已经被发展成了复杂疾病以及常见疾病的研究模型。尽管大鼠主要被用作生理学和神经生物学的模型，但是基因组学和遗传学研究中使用大鼠的次数在稳步地增加。联合应用遗传学和基因组学方法，在揭示人类功能基因组的同时，将大鼠自身的基因组计划也纳入这一过程是合乎逻辑的，也是一个能够加速完善保健措施的必然需求。

# 资源

## 遗传学和基因组学数据

1987年已经确定了10个大鼠连锁组和4个命名的染色体，构建了39种表型（表皮颜色、眼睛颜色、生长、肿瘤、牙齿等）以及33个电泳及表皮颜色标记（Robinson 1987）。在人类基因组的遗传定位改革中最主要的贡献是Weber和May使用了简单序列长度多态性（SSLP），另一个名称为CA重复和微卫星（Weber and May 1989）。同样，在大鼠中也使用了这个新的遗传学标记物。从此以后，大鼠基因组工程已经有巨大而丰富的基因组学来源，包括遗传学图谱、辐射杂交（RH）细胞系、相关的RH图谱（5000个以上的遗传学标记、19 500个基因和已定位的EST）、大于683 500个成簇出现在40 000多个UniGenes中的EST（还有很多正在产生的）的cDNA文库、10 033个以上的遗传学标记、已经发表的基于自交BN/NHsdMcwi（Brown Norway）品系的基因组序列草稿（～6.8×）（Gibbs et al. 2004）。新的测序技术将全基因组鸟枪法（WGS）和细菌人工染色体（BAC）测序技术结合起来，测出了大鼠基因组中近90%的序列。此外，Celera已经公布了Sprague-Dawley大鼠的序列草稿（1.5×；Kaiser 2005）。测出的大鼠基因组序列估计大小为2.75Gb，分布于22条染色体中的21条（Y染色体测序还没完成），预测这些序列编码大约20 973个基因，产生28 516个转录本和205 623个外显子（Gibbs et al. 2004）。基因和转录本的精确数量会在几年中得到确认，但是现在的大批数据也是可以使用的。由于大鼠测序项目的成功以及大鼠功能基因组学的价值，Mammalian Gene Collection（Gerhard et al. 2004）（全长cDNA项目）决定对先前测过的BN系大鼠的6000个全长基因进行测序，还包括完成了的4500非冗余基因。这些资源的大部分在NCBI、大鼠基因组数据库（RGD）、RatMap、UCSC、Ensembl和其他基因组数据库中发表过（表29-1）。

**表 29-1　主要大鼠资源列表**

| 数据库 | 数据类型 | 参考文献 | URL |
|---|---|---|---|
| Rat Genome Database(RGD) | 几种 | Twigger et al. 2005 | http://rgd.mcw.edu |
| RatMap | 几种 | Petersen et al. 2005 | http://ratmap.gen.gu.se |
| NCBI Rat Genome Resources | 几种 | | http://www.ncbi.nlm.nih.gov/genome/guide/rat/index.html |
| UCSC Rat Browser | 几种 | Karolchik et al. 2003 | http://genome.brc.mcw.edu/ |
| Ensembl/Rat Browser | 几种 | Hubbard et al. 2005 | http://www.ensembl.org/Rattus_norvegicus/index.html |

续表

| 数据库 | 数据类型 | 参考文献 | URL |
| --- | --- | --- | --- |
| Baylor College of Medicine Rat Resources | 序列，重叠群(contig)，拼接(assembly) | Gibbs et al. 2004 | http://www.hgsc.bcm.tmc.edu/projects/rat/ |
| Rat EST Project at the University of Iowa | 大鼠EST | Scheetz et al. 2004 | http://ratest.uiowa.edu/ |
| PhysGen Program for Genomics Applications (PGA) | 品系，表型，基因型 | Jacob and Kwitek 2002 | http://pga.mcw.edu |
| National Bio Resource Project Japan | 品系，表型，基因型 | Mashimo et al. 2005 | http://www.anim.med.kyoto-u.ac.jp/nbr/ |
| TIGR Program for Genomics Applications (TREX) | 微阵列 | | http://pga.tigr.org/ |
| NIAMS: ARB Rat Genetic Database | 品系，图谱，标记 | Dracheva et al. 2000 | http://www.niams.nih.gov/rtbc/ratgbase/ |
| Wellcome Trust Centre: Rat Mapping Resources | 图谱，标记 | Wilder et al. 2004 | http://www.well.ox.ac.uk/rat_mapping_resources/ |
| RRRC: Rat Resource and Research Center | 品系 | | http://www.nrrrc.missouri.edu/ |
| TIGR Gene Index | 基因 | Lee et al. 2005 | http://www.tigr.org/tigr-scripts/tgi/T_index.cgi? |

来自大鼠基因组序列的数据为研究人员提供了大鼠基因内容的准确知识，对于生物医学的发展具有重要意义。数据还提高了物理图谱和遗传图谱的分辨率，因为染色体上的位置信息不再依赖于重组率和统计分析。然而，还必须注意，大鼠的序列只绘成了草图，来自其他线索的证据还需要进一步获得，也就是说，遗传连锁分析或其他形式的基因定位分析还必须继续完善，使得所研究的基因组内的局部区域信息能够正确地拼接起来。目前，基因组工具箱已经近乎完善，这在本章中有论述，它对利用大鼠进行研究具有重要的影响。

## 表型数据

许多大鼠品系是为了研究多因子疾病（多基因控制加上环境影响的疾病）而选育的，然后培育成同基因型的类型。图 29-1 显示了大鼠品系之间的种系发生关系。现在，在 RDG 中已经有 1015 种大鼠品系，超过 50%是复杂性状的近郊系（538 个品系）。根据 RGD 的品系疾病和表型存在原理确定，在这些品系内有 168 种疾病和 393 种表现型。这些疾病包括：关节炎、癌症、高血压、多发性硬化症（MS）和癫痫发作。在某些研究实例中，已经培育了多种近郊系来研究单一的多因子疾病。例如，5 种不同的大鼠品系（BUF、DA、F344、LEW 和 PVG）都具有 MS 高患病风险。通过这些疾病品系中的两个（DA 和 LEW）同抗病对照品系之间的杂交结果，鉴定出了 18 个数量性状位点（QTL），它们参与了脑脊髓炎的控制，称为 MS 的动物模型（Dahlman et al. 1999a，b；Roth et al. 1999；Bersteinsdottir et al. 2000）。根据在多种品系中同样性状的重叠 QTL 置信度区间，可以鉴定疾病品系之间共有的单体型，这可以促进疾病的等位基因的位置克隆。

随着近交品系的发展，疾病相关的多个基因可随之被确定，因而可以在一个近交品种中构建多个疾病模型。然而，其中的一些性状可能仍然没有被鉴定出来。由于这个原因，需要对品系在表型水平和基因组水平上进行更好的特征鉴定。有很大的研究力量都集中在产生大鼠的表现型组。Mashimo 等［来自国立生物资源工程大鼠项目（National Bio Resource Project for Rat，NBRP）］鉴定了 54 个近交大鼠的 109 个性状（Mashimo

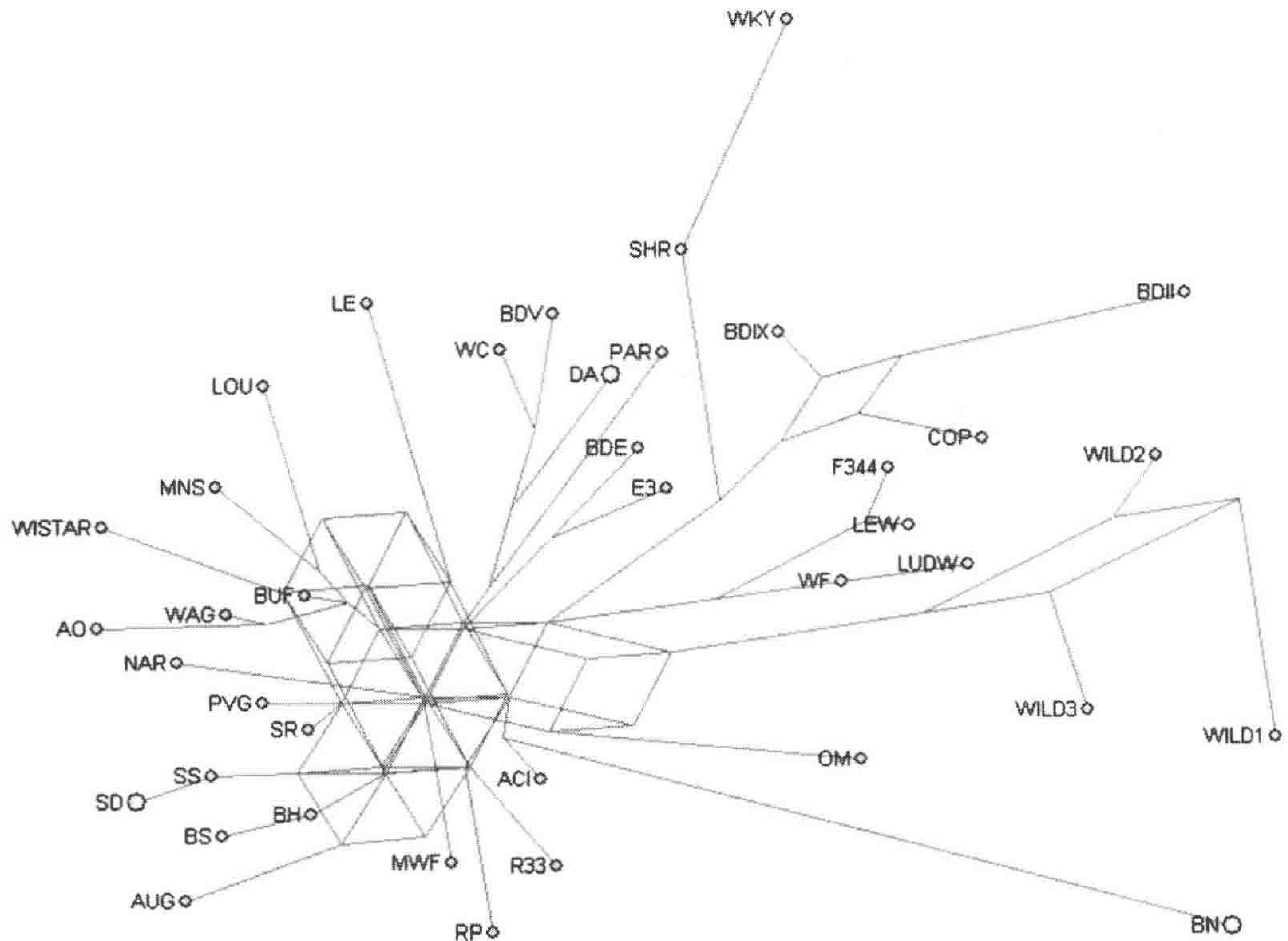



图 29-1　大鼠品系间的种系发生关系。大鼠品系关系体现为网状的结构。因为许多品系是高度相关的，其中具有潜在的复杂基因流动，祖先的节点并不能很肯定地确立。这样，数据与传统的种系发生树还是不会吻合得太好的。此图解的依据是来自 36 种普遍应用的大鼠近交品系和 3 个野生大鼠个体的 861 个 SNP 标记。末端节点（黄色点）代表的是品系。某些末端节点是两倍大小的，意味着它是依据两个样本的数据确定的。线条相遇而交汇的节点代表可能的祖先。(转引自 Smits et al.　2005 [© Bio Med Central Ltd. ])(见图版)

et al.　2005)。PhysGen (http：//pga. mcw. edu) 给出 11 个不同的品系 (9 近交系和 2 个远交系) 的 280 个不同性状，还鉴定了两条染色体的替代品样组 (44 个品系，其来源是序列已知的 BN、FHH 和 SS 高血压品系) (Jacob and Kwitek　2002；Kwitek et al.　2006)。重要的是，所有这些实验用的都是同样的方法学。

## 方法和工具

复杂疾病的新型动物模型可以将表型定位到基因组上来促进基因的鉴定，采取的手段是缩短染色体上与表型连锁的区域，在均匀的遗传学背景下确定疾病基因位点的功能。这里，我们描述了随着遗传连锁及 QTL 的研究，新型的和“设计的”大鼠模型的发展。

## QTL 定位

QTL 定位是一种已经研究清楚的方法，通过鉴定影响复杂表型的基因所存在的染色体区域，它可以将大鼠的生物学信息转化为基因组序列。尽管一个 QTL 是一个相当大的遗传位点，但其区域中包含的基因是性状变异的一个方面，使得基因组可以用生理学来解释。重要的是，多数大鼠的研究方法反映的是临床表型，一些比较定位研究已经确定了大鼠和人的常见表型通常都定位在保守的基因组区域（下文有细节内容）。QTL 定位的最终目标是使用位置克隆手段识别影响复杂表型和疾病的基因，并且能在生理学和病理生理学方面对它们有更好的认识。

到目前为止，随着生理学和病理生理学性状的 1000 个 QTL 的报道，已经发表了 536 篇 QTL 内容的文章，其中包括血压（Rapp 2000）、糖尿病（Jacob et al. 1992；Galli et al. 1996；Prarenec et al. 1996）、心血管疾病（Stoll et al. 2001；Moreno et al. 2003）、脑中风、酒精偏好程度、行为调节和焦虑、肥胖加速、关节炎、铜代谢、脑垂体肿瘤、好氧能力、化学致癌的遗传学背景的调查。最近 3 年，大多数 QTL 已经定位，主要得益于高通量基因分型技术的优势以及遗传学修饰品系的加速发展。

基于同源导入系（同类系，congenic line）的发展确定基因位点之后通常就进行 QTL 定位，定位是为了在其他已经定位的 QTL 存在时评价 QTL，这可以作为位置克隆的一步（Flint et al. 2005）。到目前为止，在大鼠中定位的 118 个 QTL 已经通过同类系确定了，它们中的大部分已经将关键基因组区间缩短到几个候选基因。这些同类系（59 个品系）的 50%被开发出来研究血压控制以及非胰岛素依赖的 2 型糖尿病品系（29 各品系）。在进行 QTL 定位的 118 个同类系，在 2002 年开始已经发表了 61 个。因而在大鼠中基因发现的频率有望能跟着这个趋势加快，而且这个趋势反映了大鼠基因序列的可用性、使用高通量序列研究序列变异的可执行性、基因鉴定的微阵列技术、途径分析以及顺式和反式调节元件的定位。这些研究策略将会大大加快基因的鉴定过程，而这些基因是隐藏在能够反映复杂疾病和表型的数百个 QTL 中。

## 比较基因定位

大鼠基因组项目的主要动力是用深厚的生物学历史解释人类常见复杂疾病的基因序列（Jacob and Kwitek 2002）。对于大鼠进行的多数研究最终都变成了翻译水平上的研究，目的是通过了解常见疾病的发病途径中的关键遗传学和生理学因子来提高人类健康水平。在不同组织中，可以通过研究基因组中进化保守的区域，对引发疾病的基因和片段进行定位（Brudno et al. 2004；Gibbs et al. 2004；Wilder et al. 2004）。目前，已经在人类疾病研究中获得了一些很好的结果。

大鼠和其他物种的基因组序列可以在核酸水平上进行比较基因组分析，这种方法要优于基于正向同源基因低分辨率排序的保守同线性鉴定。比较分析是基于一种假说：有功能的重要序列在物种中是保守的。2000 年，Stoll 等报道了大鼠 QTL 能够用来预测人类中类似 QTL 的定位情况（Stoll et al. 2000）。从此，大量的研究开始证明人类、小鼠和大鼠三个物种中进化上保守的区域（Stoll et al. 2000；Sugiyama et al. 2001；

Jacob and Kwitek 2002；Korstanje and DiPetrillo 2004)。在接下来的5年中，通过比较交叉物种找出常见复杂疾病起因的研究越来越多（Glazier et al. 2002；Korstanje and Paigen 2002)。大约100篇文章中报道了大鼠和人类相同保守区域中一种特殊的疾病性状图谱，说明了大鼠基因组工程的短期价值。然而，我们必须认识到，QTL包含的是大量的多因子性状，如行为和代谢综合征。因此，重叠的QTLs有时候仅是一个机会事件。为了说明这个问题，可以将大鼠广泛地用于评价基因的亚型，以期更好地用中间表型来匹配QTL。同时，根据大鼠SNP图谱的要求，精细分辨率定位分析可以减小QTL的大小。最终，从其他物种中（如狗、牛或其他模式生物）得来的比较定位数据可用来确定保守的QTL。具有定位数据和生物学数据的基因组序列整合后，附加到这些图谱上，再加上基因产物综合目录的创新和说明，诸如此类的分析结果都能够增加这种比较研究的使用价值，也会加大大鼠在翻译水平上研究的影响。

## “设计者”品系

### 同类系

为了验证基因组区域功能的重要性（最初是通过遗传学连锁分析鉴定的），Jackson实验室的诺贝尔获奖者Snell博士（Snell 1948）最早使用同类系技术研究小鼠中的MHC。在随后的60年里，这个技术仍然是研究基因的常用方法。自从发现同类系只与它们的背景品系在很短的染色体片段上不同之后，调查基因位点的表型效应就成为了可能，这些被调查的位点远离原始的遗传学背景下其他基因位点引起的其他效应（图29-2)。对全基因组的基因分型和培育者的选择，可以促进同类系的发展，其中除了包含供体品系的目标区域外，培育者还拥有更大的来自受体品系全基因组的等位基因比例。全基因组的辅助标记过程选择，也称为“快速同类系”(speed congenics)，能够减少一半的培养时间（Visscher 1999)。通过这个技术，同类系的产生从4～5年减少到了2～3年。复杂疾病中多个QTL决定一个性状，所以有时候为了确定一个起因基因位点，二倍体和三倍体同类系都是必要的。这么多的同类系都是特别构建的，接着它们可以组装成多重同类系（multi-congenics)。

### consomic系

consomic系是一种全染色体组来自一个品系的大鼠品系，通过与同类系相似的一种方法被转移到另一个基因组背景品系中。Wisconsin的药物学院已经组装了两个完整的consomic系，而作为供体品系的BN系序列已经知道了。在这些consomic系中，来自BN/NHsdMcwi系大鼠的一条染色体被替代到SS/JrHsdMcwi系和FHH/EurMcwi系大鼠中，而且一次替换一个。SS系大鼠是多种疾病的模型，如盐敏感高血压、胰岛素耐受性、高血脂、血管内皮功能失调、心脏高血压和肾小球硬化症。FHH系大鼠的模型包括的疾病是：心脏高血压、肾疾病、肺高血压、血液障碍、酒精中毒和抑郁症(Provoost 1994)。这两个consomic系包含了大鼠近50%的遗传学变异，同时也提供了研究心脏、肾、肺和血管障碍的遗传学资源的基础。鉴于能给出50%的遗传学变异

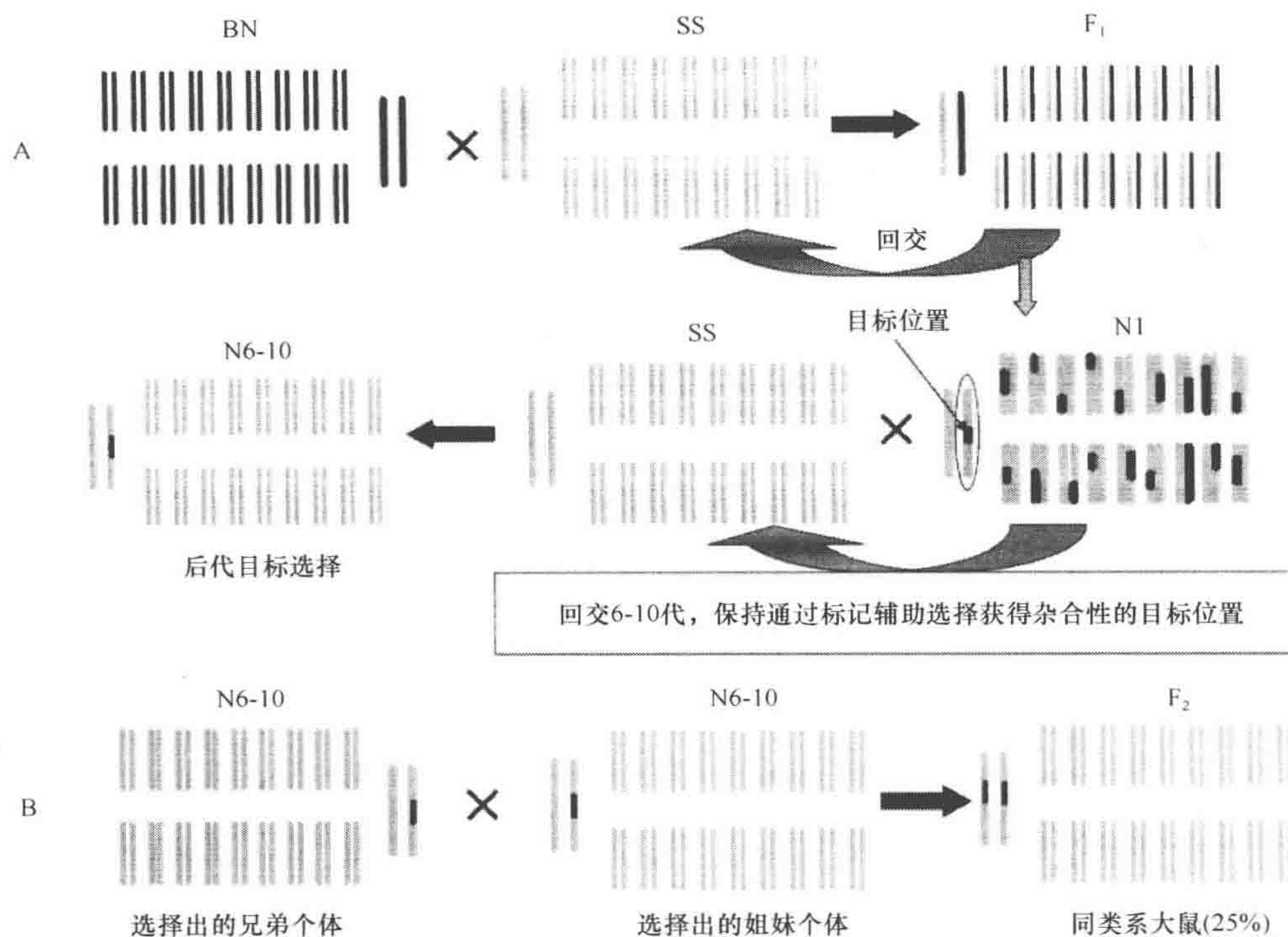



图 29-2　两种不同的大鼠品系产生同类系的原理图。A. 表示亲本品系 Brown Norway（BN）和 Dahl salt-sensitive（SS）杂交产生杂合子 $F_1$ 代。$F_1$ 与某一亲本杂交（图中是 SS）产生 N1 代。N1 代随后回交 6～10 代，通过标记辅助选择后代，这是为了代替 BN 系的选择基因组区域。B. 表示通过选择带有感兴趣表型的特殊目的片段，将雌雄小鼠交配。后代中的 25%将是这一片段的纯合子。这些小鼠可以杂交产生稳定的同类系品系。（自 Cowley 等 2004 再版）

性，假设生物学变异性的相似水平是合理的，这就使得 consomic 系被称为一种能够定位附加复杂性状的有效工具。

使用 consomic 系的一个主要优势是同类系可以快速地产生。基于 consomic 系而产生的同类系大鼠至少能通过饲养使用 3 代，同时也能进行 consomic 系和亲本系之间的杂交（图 29-3）。consomic 系大鼠能够用来评估背景效应是否对功能基因进行了修饰，也能发展成多基因模型来研究基因和基因之间的相互作用。每条染色体上的基因对观察到的性状的贡献可以通过基因分型和表达框架来评估。consomic 系和亲本系之间的比较提供了对基因组通路（成簇基因的表达方式）有价值的洞察力，而这个通路在品系之间是不同的，这些差异是如何产生的，关联到特别的病理学表型。与作为供体的亲本品系相比，consomic 系的研究优势是其所有的基因组（除了代替的染色体）在遗传上与受体品系等同，这就明显地减少了异质性。

## 重组杂交系

重组杂交系（RI）提供了通过基因组确定基因型的另一种工具。这一策略的基础

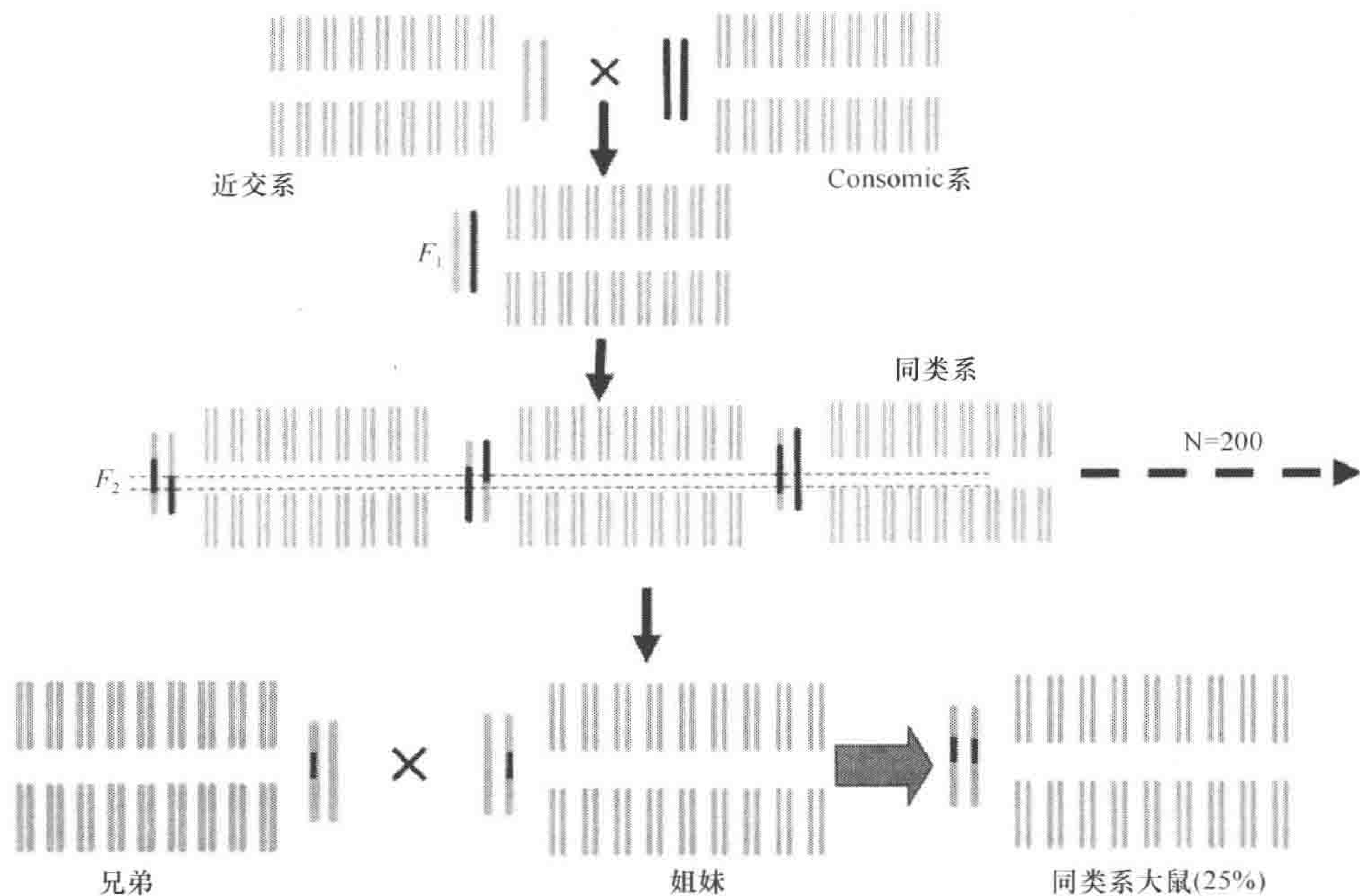



图 29-3　consomic 系产生同类系大鼠。亲本与 consomic 系杂交产生 $F_1$ 代，$F_1$ 代有相同的遗传学背景而目标染色体是杂合子。这些 $F_1$ 代互交产生 $F_2$ 代，$F_2$ 代中目的染色体将会是同类系，不过这取决于重组事件。通过基因型选择两个相似的 $F_2$ 代大鼠交配，对目的区域进行固定。（自 Cowley 等 2004 再版）

来自 $F_2$ 代的一系列杂交系的产生（Pravence et al.　1996）。包含多个 QTL 的品系产生可以进行基因相互作用的分析以及弱基因位点的检测。然而，每种品系独特的基因组背景阻止了它们在快速产生同类系动物中的使用。

最大的啮齿类重组杂交品系之一是互补的 HXB/BXH 重组杂交系，源自自发性高血压（SHR）和 BN 大鼠品系（Pravenec et al.　1989，1999）。这些品系是对心血管和代谢表型进行遗传学分析的良好资源。RI 品系的使用已经促进了包括血压（Pravenec et al.　1995）、生殖性状、代谢性状、行为和癌症易患性在内的一些性状的定位。尽管一些 RI 基因型的特征还不像 HXB/BXH 一样全面，但也在发展中，如 LEXF 和 SWXJ 系列。

## 异质品系

异质品系（HS）是通过 8 种杂交系的互交后又连续远交产生的（Hansen and Spuhler　1984）。尽管这一方法产生于人类和大鼠基因组计划之前，但是大鼠基因组计划的资源使得大鼠品系集合异常大。HS 系后代的染色体代表了小鼠的随机镶嵌，而这些打造好的大鼠在重组事件中都有一个接近 1cM 的平均距离。这种高度的重组使得精细的 QTL 定位到亚 cM，先前识别的单个 QTL 变成了多种 QTL（如 Ariyarajah et al.　2004；Stylianou et al.　2004）。HS 系群体来自于 1984 年对 NIH 系进行的酒精研究（Pandey et al.　2002）。在 HS 系大鼠中很多性状的 QTL 都精细定位到亚 cM 的距

离，如焦虑症（Mott et al. 2000）、酒精引起的运动活力（Demarest et al. 2001）和条件性恐惧（Mott et al. 2000）。现在的研究都是在 HS 系大鼠中检测多重性状的 QTL，如行为和糖尿病性状。

## 位置克隆基因（positionally cloned gene）

在过去的 2 年中，大鼠中传统的位置克隆在鉴定疾病基因方面已经取得了不错的成果。通过位置克隆许多基因被鉴定出来，同时大鼠基因组资源也在增加。已经鉴定出的基因有：肿瘤基因（*BHD*，*Tsc2*）、1 型糖尿病基因（*Gimap 5*，*Cblb*）、神经功能失调基因（*Cct4*，*Reln*，*Unc5h3*）、关节炎基因（*Ncf1*）、肾疾病基因（*Pkhdl*，*Rab38*）、肾小球肾炎基因（*Fcgr3*）、出血障碍基因（*Rab38*，*VKOR*）、视网膜退变基因（*Mertk*）、hypotrichosis 基因（*Dsg4*，*Whn*）。它们中的很多基因都是从符合孟德尔遗传规律的自发突变体上克隆出来的，如 PCK 大鼠的 *Pkdh1* 突变体会引发常染色体隐形多囊性肾病。然而鉴定出的复杂疾病基因数量在增加。

基因组学中最具挑战性的任务是预测基因功能和研究基因之间的相互作用，也就是功能基因组学。DNA 微阵列可以用于研究疾病的发病机制中的基因和通路，以及对生理学应激物、药物、环境刺激的应答反应的基因和通路。大鼠微阵列研究已经和其他的遗传学方法联合使用来加速各种表型的基因研究，如 QTL 分析、同类系定位、转基因技术。基因 *Cd36* 是最早联合使用几种方法在大鼠中克隆的复杂性状的基因之一，方法是将 QTL 引入产生的同类系中，找到其相对于亲本的表达模式。更新一些的研究是在一系列的 HXB/BXH 重组杂交（RI）系中研究基因的表达，以此在大鼠基因组中鉴定表达 QTL（eQTL）。与先前鉴定的代谢综合征 QTL 重叠的 eQTL 提供了近 76 个有待评价的候选基因。

## 靶位确认（target validation）

一旦一个基因进行了位点克隆或有迹象表明其有某种因果关系，那么其因果关系就必须确认。靶位确认（target validation）就是表达这一基因发现阶段的术语。证实某个基因是有因果关系的黄金标准是在特定等位基因上实行敲除或突变以观察相应性状的改变，也可以通过用“正常”等位基因代替缺陷的基因来验证这样取代造成的表型变化。尽管这些措施可以给出结论性的证据，但要在全部 30 000 个以上基因中应用，是不可能的，因为这项工程过于繁烦也过于昂贵。对于大鼠的基因组，目前已经另辟蹊径，即转基因拯救（transgenic rescue）——通过转基因技术使得表型正常化，特别是当性状表现为隐性遗传模式时（pravence et al. 2001；Jacob and kwitek 2002）。最近，通过核转移技术已经完成了成体可育大鼠的克隆工作（Zhou et al. 2003），这为诸如基因敲除（knock-out）之类的靶向基因操作开启了一扇门。但是，这种手段的效率还太低，还不能作为普遍的手段，所以，距离将其常规应用还有一段时间。

## 转基因大鼠

在 15 年以上的时间内，遗传学研究是从两个方向揭示基因功能的：从遗传学定位

到基因鉴定的位置克隆和转基因（随机地插入基因、敲除基因、基因敲入和有条件的基因敲除）。1990 年，开创了传统的原核注射的大鼠转基因方法（Hammer et al. 1990）。然而，由于缺乏存活的大鼠 E S 细胞株，传统的基因敲除和基因敲入技术是不能用的，这就限制了大鼠的基因研究。但是，200 多种转基因小鼠已经产生了。

和小鼠一样，产生转基因大鼠的主要目的是研究感兴趣的特定基因。现在，改变特殊基因的表达和使用大鼠作为其他物种基因表达的替代宿主一样都在相对地向前发展。通过使用人类基因，许多转基因大鼠已经被人源化了，这为人类遗传学连锁研究提供了一个桥梁，也提供了有特定生理特征的（突变）基因在功能上的关联。例如，人源化大鼠可以用来剖析复杂疾病，如心肌肥大、终末器官损伤、高血压等。这些例子证实人类疾病的大鼠模型是有价值的，我们希望通过序列信息和转基因方法能证实其他疾病的病因。进一步讲，表达人类基因的转基因大鼠能研究体内疾病的纵向进程，监测长期治疗的效果，包括细胞植入、病因的反义方法。最终，在转基因治疗中通过位置克隆的基因需要验证，这就会进一步促进转基因技术的使用。

### *N*-乙基-*N*-亚硝基脲的诱变效应

许多年来，遗传学团体以自发的突变作为大鼠模型的来源，这是使用大鼠进行结构功能研究最主要的限制因素，尤其是对疾病基因的验证。而使用化学诱变剂产生的突变体能够作为替代物，如 N-乙基-N-亚硝基脲（ENU）。ENU 诱变作用已经使用了很多年，但是近 10 年来其使用呈上升趋势（Guenet 2004）。典型的研究手段是通过 ENU 创造雄性动物，引起精原干细胞（缺失主要功能）的突变，与未处理的雌性繁殖，以基因型效应筛选后代。随着 ENU 技术的改进，在大范围的研究中表型的确定逐渐成为限制速率的一步。对小鼠系统的综合基因功能分析越来越强调大范围的 ENU 筛选（Nolan et al. 2000）。突变项目是表型驱动的，采用合适的方法筛选突变动物以确定新的突变表型，尤其是那些人类模式疾病的表型，之后就能对引起疾病的基因进行定位和分离。然而，大范围表型筛选的费用是非常昂贵的，每日都会消耗大量的动物。如果是这样，在大鼠中就不能使用这种筛选进行相似的尝试。

2003 年，有两个研究小组报道了联合使用 ENU 和基因筛选产生基因敲除大鼠。主要的不同在于目的基因的筛选，并且仅仅保存拥有特定基因变异的动物（Zan et al. 2003）。通过这条途径，尽管基因筛选是有消耗的，但是每日的消耗达到了最少。这个手段被用来敲除大鼠的某些基因（Smits et al. 2006）。另外一个优势就是 ENU 的诱变效应可以用来选择品系（决定了合适的 ENU 的剂量之后）。这就预防了与基因组背景效应有关的问题，而且有限 ES 细胞株产生的基因敲除效应相对来说是普遍的。直到有了大鼠的 ES 细胞后，此方法会更受欢迎，并且当基因组背景影响了表型时也仍然是一种备选方案。目前 RGD 上登记了 26 种 ENU 诱导的突变系。

## 基因变异

尽管 QTL 定位提供了有价值的基因组信息，仍然需要对影响复杂性状基因的鉴定

工具进行改进。目前，QTL 能被局部定位于 2～10cM 的特殊基因间距，但是进一步的研究却是有难度的。SNP 和单倍体图谱是缩短 QTL 间距大小的有效工具。SNP 的发现和基于 SNP 的单倍体图谱可以通过两个普通途径减少这个间距。第一，表型数据和祖先序列之间关联的进步，后者来源于许多现存的杂交系。这个方法可以快速鉴定短的基因组片段，而这个片段是最有可能含有决定性状基因。第二，鉴定在简单互交和回交实验中所用大鼠的共有片段。

## SNP 和单体型

DNA 水平上的变异是作用于种群中个体表型多样性的主要因子。最常见的遗传学变异类型是 SNP。尽管绝大多数 SNP 都是没有功能的，但是其余的则能影响染色体组装、基因表达或者蛋白质功能。SNP 和它们各自的区间（等位基因）在基因组和种群中不是随机地分布。重组和突变事件以及选择过程和种群历史导致在基因组中产生了像块状的结构。SNP 的等位基因发生的常见重组（也称单体）鉴定了这些结构特征。通过单体型区段的个体多态性的优势效应，可在种群中选择特殊的单体型。

不像远亲后代种群（人类）的个体之间的变异，在很多物种中都能观察到上述这种镶嵌性的分区方式，并且在通常使用的小鼠杂交品系中有相对少的变异（Wiltshire et al. 2003）。此外，对全基因组 SNP 数据（所取的样分辨率相对低）进行成对的比较，发现变异发生在有极高（每 10kb 有多于 40 个）或者极低（每 10kb 小于 1 个）水平多态性的镶嵌性区域。这些区域大小为 10～120Mb，被认为是最近遗传学的瓶颈所产生的结果（Wiltshire et al. 2003）。由于在中国和日本，野生小鼠和大鼠已经被驯化了，所以实验室常用的品系来自很小的选择范围。

单体型信息可以通过两种方式加速实验室动物的定位和克隆过程。第一，认为杂交种个体之间没有遗传学变异，因此要确定单体型区段结构。作为拥有共同祖先的结果，等位基因相似度的方式和品系中的不同（品系分布方式或者 SDP）能够被每个可变基因位点所识别（Grupe et al. 2001）。理论上分析，可以通过关联的表型和基因型 SDP 来对突变体作图，因为常见的遗传多态性更容易引发常见的表型性状，在不同品系中获得的独立的新突变体则不易引发表型性状。最近一些报道尽管给出了方法的一般适用性，但也指出了此方法的原理证据（Wang et al. 2004）。

第二，基因分型在每一个单体型区段（标记 SNP）选择的 SNP 是一个有限的数字，这就足够接近所有在区段中的其他多态性（Sebastiant et al. 2003）。这个方法减少了进行基因分型的 SNP 数目，但是却存在一个缺点，目的多态性的基因组区域不能被进一步缩短成为最短的常见单体型区段结构或者 SDP。由于品系有共同的祖先，这些区域能够扩展到与遗传的片段交叉，并且能够包含数百种被表征过的和未被表征的多态性。引进一个额外的仔细挑选过的品系组（如对它们品系发生之间的关系加以重视）希望被用来增加区段的数量和提高用于基因鉴定所需要的分辨率。

## 大鼠 SNP 数据

目前关于大鼠基因组变异和 SDP 的信息是基于极少的 SNP 数据。dbSNP（126 个）

包含了 43 229 个大鼠 RefSNP。这些数据既包括了 cDNA 和非转录基因组区域的 SNP，还提供了有关大鼠基因组 SNP 预期频率的重要信息。Zimdahl 等（2004）对 SHRSP、BN、WKY、SD 系的 cDNA 进行测序，在品系之间比较鉴定了 12 395 个多态性位点。通过对 BN 系和其他品系的比较，判断发现的比例是 cDNA 上每 1100bp 有一个 SNP。Smits 等（2004b）在 96 个品系中筛选了 55 个基因，找到了 103 个 SNP。仅仅考虑基因内 SNP，将相近的多态性组成一个单一的 SNP，研究者推算出每 367bp 出现一个 SNP，明显高于先前基于大量品系的研究结果。

Guryev 等（2004）使用发表的 WGS 可用序列、EST、mRNA 数据，进行电子杂交来鉴定基因编码区域的 33 305 个高质量的候选 SNP。使用有限的大鼠组进行实验验证了 471 个候选 SNP，确认率大约为 50%。尽管在 Sprague-Dawley（EST 数据）品系和 Brown Norway（WGS 数据）品系鉴定了大多数 SNP，但是验证了的变异中 66%都是在不同大鼠品系中常见的（小的等位基因频率大于 20%）。基于被证实的数据，得到的 SNP 的频率是 226 bp 一个。由于半数多态性现象都是基因内的，所以得到的这个频率比先前的数字要高。

因为大多数 SNP 的发现方法都以品系的选择为基础，Smits 等（2005）使用鸟枪测序法对野生大鼠品系测序，与 BN 基因组序列相比，在 814kbp 长的序列中发现了 485 个 SNP。有趣的是，对 36 个普遍使用的杂交大鼠进行基因分型，结果表明这些等位基因的 84%代表了一系列实验室品系的多态性。通过双脱氧序列进行基因分型，发现了额外的 358 个 SNP（这些序列的数据可以使用多种软件包分析）。基于基因分型的结果，BN 系和野生大鼠之间的 SNP 是 190bp 产生一个 SNP。在 36 个大鼠品系（包括 BN）中的 SNP 比例是 158bp 产生一个。

最终，Celera 在 Sprague-Dawley 系小鼠中用基因组鸟枪法测序（1.5×覆盖度）。初步的分析表明通过与 BN 系的比较，在这些数据中发现了 2 万～4 万个 SNP，结果是创建了两个品系（400～800bp 有一个 SNP）的平均 SNP 频率。

## 大鼠单体型数据

更广泛的是，Yalcin 等（2004a）、Frazer 等（2004b）以及近期的 Gutyev 等（2006）使用的精细扫描分析已经注意到啮齿动物单体型区段结果和其在大小鼠遗传学研究的效用。Frazer 和其同事在 13 个常用的杂交小鼠系和 2 个野生型杂交小鼠系中，分析了 5 个基因组间隔，总的大小为 4.6Mb（小鼠基因组的 0.17%）。使用高密度的寡核苷酸阵列，他们对这些区域大小为 3Mb 的非重复序列进行了重新测序，并且发现了 18 366 个 SNP。然而，在 13 个常用杂交小鼠系中仅有 4065 个 SNP 是多态性的。对隐藏的 Markov 模型进行分析，设想了一个存在两个状态的模型，这个模型拟合了高低两种 SNP 的频率并且集合了 10 000 个核苷酸（Wade et al. 2002），而前面的 Marker 模型大约有 50 个单体型区段，大小为 12～608kb。多数情况表示不同遗传亚种 *Mus musculus musculus* 和 *Mus musculus domesticus* 的分布。平均来说，区分一个特定的单体型区段只需要 1～3 个 SNP。将这些结果外推到全部的 2.5Gb 的小鼠基因组中，表明了一个已评估过的 50 000bp 的标记 SNP 对于定义杂交品系中任何片段之间系统发育关系的需要。

尽管这些结果与先前的数据是一致的（Wade et al. 2002)，但这项研究也表明了至少在12个杂交系中序列的细节信息对于捕获变异位点的大多数（>95%）是必要的，而那些位点存在于当前普遍使用的杂交小鼠品系中。这意味着，至少12个杂交大鼠系的重新测序是可靠的全基因组单体型区段鉴定和后续的SNP标记鉴定所必需的。此外，通过微阵列重新测序、品系选择和计算评估得到的假的阴性比例（>50%）仍然可以作为与提出的单体型区段相关联的变量。

Yalcin和其同事分析了包含有影响小鼠焦虑的QTL所在的一个连续的大小为4.8Mb的区域（Yalcin et al. 2004b)。这是对QTL定位中，将单体型区段和SDP联合使用进行有效性的检测。通过对8个杂交小鼠系的8kb间隔进行1～2kb的重新测序，获得了高分辨率的SNP数据。总的来说，取0.58Mb序列作为样品，结果鉴定了1720个变异体，其中1325个是SNP。SNP密度的分析证实了具有特别低或特别高的SNP密度的片段的镶嵌型。具有中等SNP密度的片段也被观察了。然而，从现时的SNP抽样密度来看，还不清楚是否这些片段包含了两种较小片段的混合物。

在近期的一项研究中，Guryer等（2006）比较了大鼠、小鼠和人类的单体型结构，该结构在大鼠的1号染色体上占有5Mb的区域。他们还在小鼠和人类中比较了其保守的同线区域。结果显示单体型结构在哺乳动物中是保守存在的，并且在基因区域存在显著的保守现象，这暗示了存在着一种进化选择的过程，该过程驱动了长距离等位结合的保守存在（图29-4)。实际上，在基因组水平上，人类HapMap数据的以基因为中心的分析揭示了在基因区和它们上游的调控区中，相同空间的多态性位点比在非基因区中的位点从遗传学上来看更能紧密地连接。这些发现可能会使基于表型特征基础上的因果多态现象的鉴定变得更困难，因为在单体型结构保守的区域内，紧密连接的多态现象的结合（而不仅仅是单个偶然出现的SNP）可能影响表型的不同。另外，单体型结构在进化上的保守可实际被用于鉴定和定义功能上重要的基因组区域。图29-5将大鼠第10号染色体和人类的序列进行了比较。

## 发展中的SNP计划

基于单体型的大鼠遗传学正在经历着里程碑式的发展过程。为了全面了解大鼠基因组序列的作用，研究者提出并正在进行着两个行动倡议来分析祖先留下的这些片段，这些片段组成了现在最常用到的近交系。其中一个倡议名叫STAR，它是由欧盟建立并旨在建成一个大鼠的单体型图谱，其主要目标是在至少200个不同的近交系中鉴定出100 000个左右的SNP的基因型。这些基因型数据将对建立关联表现型和各祖先原始品系之间的联系非常有用，这就可以使对那些对应于性状的决定性区域的鉴定和高质的绘图成为可能。通过研究用于QTL图谱试验的所有重要大鼠系的内容，研究者可以通过连锁分析的方法缩短临界范围，这种方法利用的研究对象是所研究品系中的共有片段，或者，研究者也可以为杂交/回交试验选出理想的品系结合体。第二个倡议是由NIH支持的一个行动。大鼠基因组测序协会将启动一个大型SNP探索行动，该行动通过鸟枪法对8个常用的近交大鼠系进行测序，目的在于从这8个品系中找出多于280 000个的SNP。

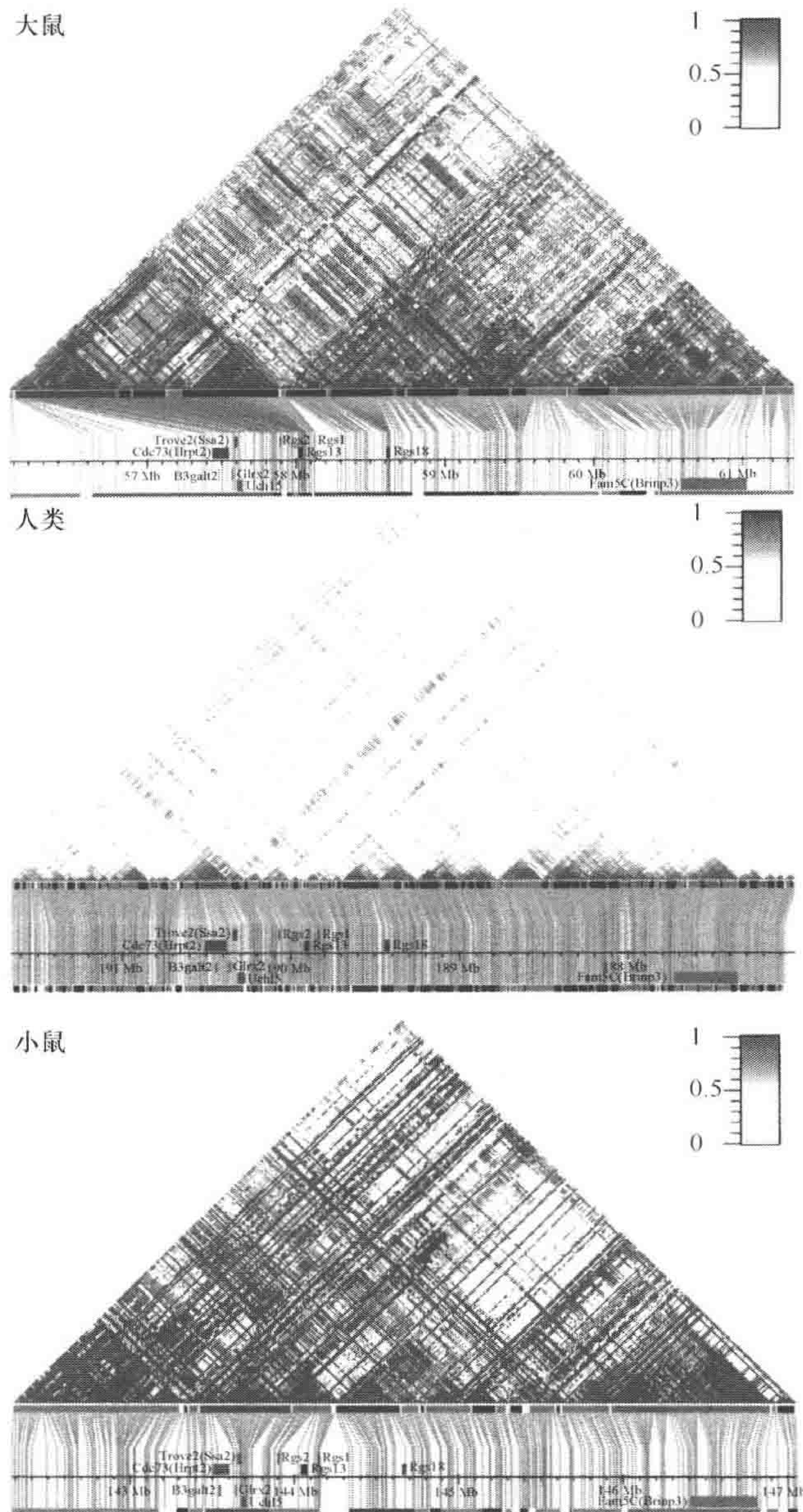



图 29-4　大鼠、人和小鼠中，LD 的大小为 5Mb 的正向同源基因组片段的模式 。图中显示了大鼠（1351 个 SNP）、人（1479 个 SNP）和小鼠（311 个 SNP）的正向同源基因组片段的 LD 图。每一个分析组（panel）都标明了以下信息：LD 图（上部），SNP 坐标中的单体型区段（中部），物理坐标上的物理图和单体型区段（下部）。单体型图谱中用梯度表示了 D' 值，而后者能够帮助单体型区段结构进行可视比较。单体型区段与应急标准同时建立，有时可以导致视觉上重组区段的分离。小鼠和大鼠杂交品系中的 LD 模式有共同的特征。两个生物体都展示了与以下基因组片段相应的增加着的 LD 可扩展区段：①5 个基因的基因簇：B3galt2、Cdc73（Hrpt2）、Glrx2、Trove2（Ssa2）和 Uchl5；②大的 Fam5C（Brinp3）基因；③Rgs18 基因旁侧的区域。尽管人类单体型区段结构特征是有许多更小的区段，而这些区段是表达高 LD 的最广泛的人类基因区域。因此对单体型区段进行扩展，包括了上面提到的含有 5 个基因的基因簇、Fam5C（Brinp3）的编码片段和 Rgs18 基因旁侧的区域。在杂交物种中保守的 3 个有特征的单体型区段已经用颜色标记了（Guryev 等 2006 年再版）

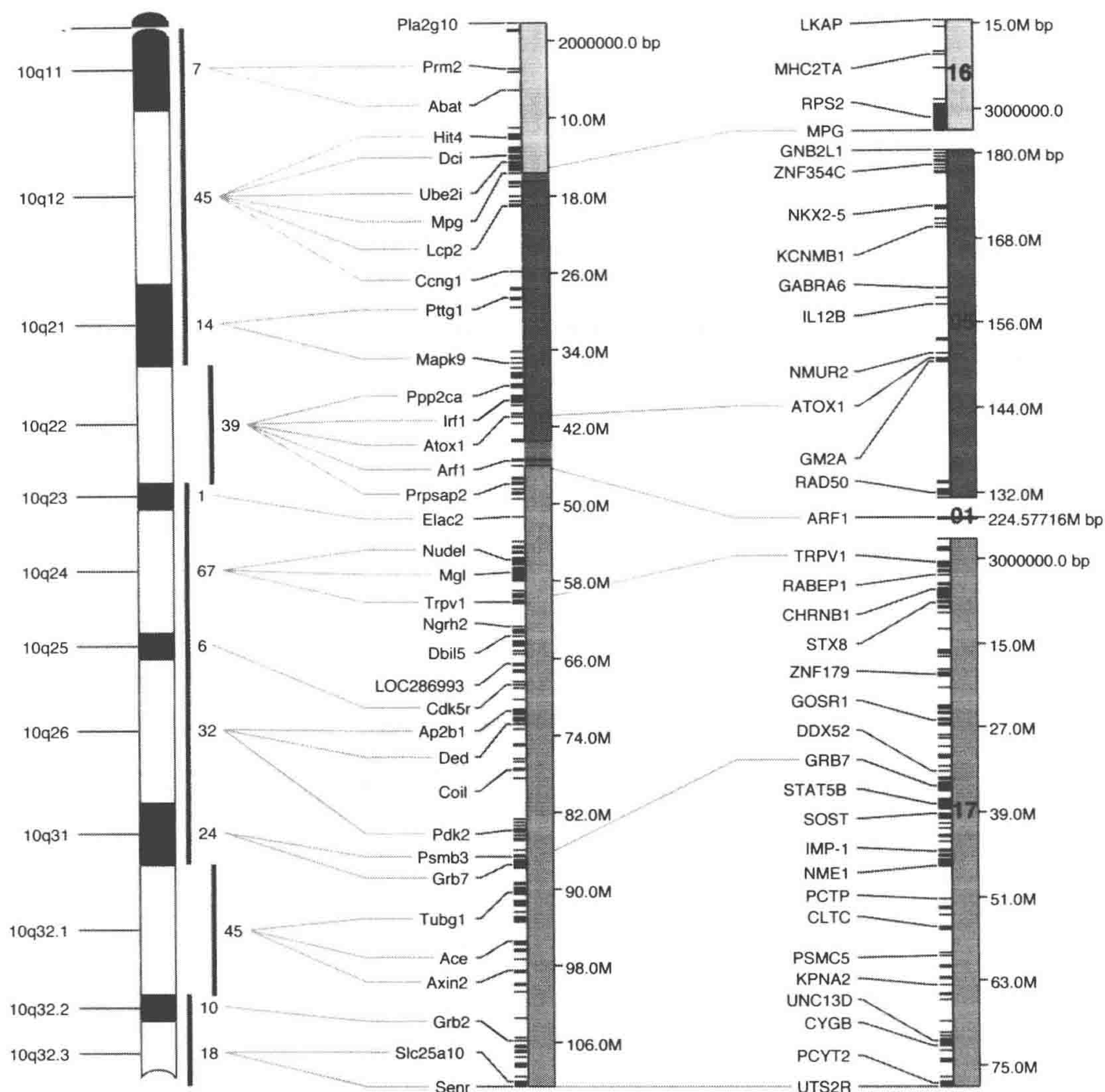



图 29-5　利用 VCMap 工具得到的大鼠 10 号染色体和人类的比较图谱。左侧 RNO10 的细胞遗传图。细胞遗传图右侧的横线表示进行基因绘图的细胞遗传的目标。数字表示的是每个目标所包含的基因数目。中间是 RNO10 的基因组图谱，用 DNA 百万碱基对（Mb）测量计数。基因符号被标在了图的左侧。有颜色的区域代表的是与人类基因组相保守的同线型。右侧是人类保守的片段，用人类染色体数标记并在每个染色体上以 Mb 测量。16 号染色体为蓝色，5 号染色体为红色，17 号为青绿色（见图版）

这种程度上的覆盖范围就可以满足绘制基因定位图谱的需要，还可以深入评估为了最大地实现致病基因的探索还需要哪些额外的资源。上述两种行动中所用品系都被选来用于涵盖实验室大鼠的已知遗传变异的“进化”范围，除了使差异实现最大化以外，这些品系也包括了那些大量 QTL 已经被绘图的品系或者那些已经存在了其他基因工具的品系，这里提到的基因工具包括同类系、consomics 系和重组近交系。

从任何一个单体型图谱计划中获得的 SNP 数据肯定会对很多计划都具有启发性，

其中包括不依赖于单体型功能块（haplotype-block）定义的基因绘图和克隆方法（Liao et al. 2004；Pletcher et al. 2004）。高通量 SNP 分型策略、基因表达数据图谱和生理学 QTL 的结合（Hubner et al. 2005；Malek et al. 2006）、不断增效的候选区域的重新定序都可用来帮助克服很多与传统绘图方法和克隆方法相关的瓶颈问题。基因组序列和更大的 SNP 数据库的结合也将进一步完善研究者利用大鼠来研究人类疾病时所用到的工具和方法。

# 参考文献


Aitman T.J., Glazier A.M., Wallace C.A., Cooper L.D., Norsworthy P.J., Wahid F.N., Al-Majali K.M., Trembling P.M., Mann C.J., Shoulders C.C., et al. 1999. Identification of *Cd36* (*Fat*) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nat. Genet.* **21:** 76–83.

Aitman T.J., Dong R., Vyse T.J., Norsworthy P.J., Johnson M.D., Smith J., Mangion J., Roberton-Lowe C., Marshall A.J., Petretto E., et al. 2006. Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **439:** 851–855.

Ariyarajah A., Palijan A., Dutil J., Prithiviraj K., Deng Y., and Deng A.Y. 2004. Dissecting quantitative trait loci into opposite blood pressure effects on Dahl rat chromosome 8 by congenic strains. *J. Hypertens.* **22:** 1495–1502.

Bergsteinsdottir K., Yang H.T., Pettersson U., and Holmdahl R. 2000. Evidence for common autoimmune disease genes controlling onset, severity, and chronicity based on experimental models for multiple sclerosis and rheumatoid arthritis. *J. Immunol.* **164:** 1564–1568.

Bila V. and Kren V. 1996. The teratogenic action of retinoic acid in rat congenic and recombinant inbred strains. *Folia Biol.* **42:** 167–173.

Bohlender J., Ganten D., and Luft F.C. 2000. Rats transgenic for human renin and human angiotensinogen as a model for gestational hypertension. *J. Am. Soc. Nephrol.* **11:** 2056–2061.

Bonfield J.K., Rada C., and Staden R. 1998. Automated detection of point mutations using fluorescent sequence trace subtraction. *Nucleic Acids Res.* **26:** 3404–3409.

Brown S.D. and Balling R. 2001. Systematic approaches to mouse mutagenesis. *Curr. Opin. Genet. Dev.* **11:** 268–273.

Brudno M., Poliakov A., Salamov A., Cooper G.M., Sidow A., Rubin E.M., Solovyev V., Batzoglou S., and Dubchak I. 2004. Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.* **14:** 685–692.

Cardon L.R. and Abecasis G.R. 2003. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19:** 135–140.

Conti L.H., Jirout M., Breen L., Vanella J.J., Schork N.J., and Printz M.P. 2004. Identification of quantitative trait loci for anxiety and locomotion phenotypes in rat recombinant inbred strains. *Behav. Genet.* **34:** 93–103.

Cowley A.W., Jr., Roman R.J., and Jacob H.J. 2004. Application of chromosomal substitution techniques in gene-function discovery. *J. Physiol.* **554:** 46–55.

Cuppen E. 2005. Haplotype-based genetics in mice and rats. *Trends Genet.* **21:** 318–322.

Dahlman I., Jacobsson L., Glaser A., Lorentzen J.C., Andersson M., Luthman H., and Olsson T. 1999a. Genome-wide linkage analysis of chronic relapsing experimental autoimmune encephalomyelitis in the rat identifies a major susceptibility locus on chromosome 9. *J. Immunol.* **162:** 2581–2588.

Dahlman I., Wallstrom E., Weissert R., Storch M., Kornek B., Jacobsson L., Linington C., Luthman H., Lassmann H., and Olsson T. 1999b. Linkage analysis of myelin oligodendrocyte glycoprotein-induced experimental autoimmune encephalomyelitis in the rat identifies a locus controlling demyelination on chromosome 18. *Hum. Mol. Genet.* **8:** 2183–2190.

De Miglio M.R., Pascale R.M., Simile M.M., Muroni M.R., Calvisi D.F., Virdis P., Bosinco G.M., Frau M., Seddaiu M.A., Ladu S., and Feo F. 2002. Chromosome mapping of multiple loci affecting the genetic predisposition to rat liver carcinogenesis. *Cancer Res.* **62:** 4459–4463.

de Wolf I.D., Bonne A.C., Fielmich-Bouman X.M., van Oost B.A., Beynen A.C., van Zutphen L.F., and van Lith H.A. 2002. Quantitative trait loci influencing hepatic copper in rats. *Exp. Biol. Med.* **227:** 529–534.

Demarest K., Koyner J., McCaughran J., Jr., Cipp L., and Hitzemann R. 2001. Further characterization and high-resolution mapping of quantitative trait loci for ethanol-induced locomotor activity. *Behav. Genet.* **31:** 79–91.

Dracheva S.V., Remmers E.F., Chen S., Chang L., Gulko P.S., Kawahito Y., Longman R.E., Wang J., Du Y., and Shepard J. 2000. An integrated genetic linkage map with 1137 markers constructed from five F2 crosses of autoimmune disease-prone and -resistant inbred rat strains. *Genomics* **63:** 202–226.

Fernandez-Teruel A., Escorihuela R.M., Gray J.A., Aguilar R., Gil L., Gimenez-Llort L., Tobena A., Bhomra A., Nicod A., Mott R., et al. 2002. A quantitative trait locus influencing anxiety in the laboratory rat. *Genome Res.* **12:** 618–626.

Flint J. 2003. Analysis of quantitative trait loci that influence animal behavior. *J. Neurobiol.* **54:** 46–77.

Flint J., Valdar W., Shifman S., and Mott R. 2005. Strategies for mapping and cloning quantitative trait genes in rodents. *Nat. Rev. Genet.* **6:** 271–286.

Frazer K.A., Pachter L., Poliakov A., Rubin E.M., and Dubchak I. 2004a. VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **32:** W273–279.

Frazer K.A., Wade C.M., Hinds D.A., Patil N., Cox D.R., and Daly M.J. 2004b. Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. *Genome Res.* **14:** 1493–1500.

Gal A., Li Y., Thompson D.A., Weir J., Orth U., Jacobson S.G., Apfelstedt-Sylla E., and Vollrath D. 2000. Mutations in *MERTK*, the human orthologue of the RCS rat retinal dystrophy gene, cause retinitis pigmentosa. *Nat. Genet.* **26:** 270–271.

Galli J., Li L.S., Glaser A., Ostenson C.G., Jiao H., Fakhrai-Rad H., Jacob H.J., Lander E.S., and Luthman H. 1996. Genetic analysis of non-insulin dependent diabetes mellitus in the GK rat. *Nat. Genet.* **12:** 31–37.

Ganguli M., Tobian L., and Iwai J. 1979. Cardiac output and peripheral resistance in strains of rats sensitive and resistant to NaCl hypertension. *Hypertension* **1:** 3–7.

Gerhard D.S., Wagner L., Feingold E.A., Shenmen C.M., Grouse L.H., Schuler G., Klein S.L., Old S., Rasooly R., Good P., et al. 2004. The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC). *Genome Res.* **14:** 2121–2127.

Gibbs R.A., Weinstock G.M., Metzker M.L., Muzny D.M., Sodergren E.J., Scherer S., Scott G., Steffen D., Worley K.C.,

Burch P.E., et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Glazier A.M., Nadeau J.H., and Aitman T.J. 2002. Finding genes that underlie complex traits. *Science* **298:** 2345–2349.

Grupe A., Germer S., Usuka J., Aud D., Belknap J.K., Klein R.F., Ahluwalia M.K., Higuchi R., and Peltz G. 2001. In silico mapping of complex disease-related traits in mice. *Science* **292:** 1915–1918.

Guenet J.L. 2004. Chemical mutagenesis of the mouse genome: An overview. *Genetica* **122:** 9–24.

Guryev V., Berezikov E., and Cuppen E. 2005. CASCAD: A database of annotated candidate single nucleotide polymorphisms associated with expressed sequences. *BMC Genomics* **6:** 10.

Guryev V., Berezikov E., Malik R., Plasterk R.H.A., and Cuppen E. 2004. Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Res.* **14:** 1438–1443.

Guryev V., Smits B.M.G., van de Belt J., Verheul M., Hubner N., and Cuppen E. 2006. Haplotype block structure is conserved across mammals. *PLoS Genet.* **2:** e121.

Hammer R.E., Maika S.D., Richardson J.A., Tang J.P., and Taurog J.D. 1990. Spontaneous inflammatory disease in transgenic rats expressing HLA-B27 and human $\beta_2$m: An animal model of HLA-B27-associated human disorders. *Cell* **63:** 1099–1112.

Hansen C. and Spuhler K. 1984. Development of the National Institutes of Health genetically heterogeneous rat stock. *Alcohol Clin. Exp. Res.* **8:** 477–479.

Hocher B., Liefeldt L., Thone-Reineke C., Orzechowski H.D., Distler A., Bauer C., and Paul M. 1996. Characterization of the renal phenotype of transgenic rats expressing the human endothelin-2 gene. *Hypertension* **28:** 196–201.

Homberg J.R., Olivier J.D., Smits B., Mudde J., Cools A.R., Ellenbroek B.A., and Cuppen E. 2005. O9 phenotyping of the serotonin transporter knockout rat. *Behav. Pharmacol.* (suppl. 1) **16:** S21.

Hrabe de Angelis M.H., Flaswinkel H., Fuchs H., Rathkolb B., Soewarto D., Marschall S., Heffner S., Pargent W., Wuensch K., Jung M., et al. 2000. Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat. Genet.* **25:** 444–447.

Hubbard T., Andrews D., Caccamo M., Cameron G., Chen Y., Clamp M., Clarke L., Coates G., Cox T., Cunningham F., et al. 2005. Ensembl 2005. *Nucleic Acids Res.* **33:** D447–D453.

Hubner N., Wallace C.A., Zimdahl H., Petretto E., Schulz H., Maciver F., Mueller M., Hummel O., Monti J., Zidek V., et al. 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37:** 243–253.

Jacob H.J. and Kwitek A.E. 2002. Rat genetics: Attaching physiology and pharmacology to the genome. *Nat. Rev. Genet.* **3:** 33–42.

Jacob H.J., Pettersson A., Wilson D., Mao Y., Lernmark A., and Lander E.S. 1992. Genetic dissection of autoimmune type I diabetes in the BB rat. *Nat. Genet.* **2:** 56–60.

Jahoda C.A., Kljuic A., O'Shaughnessy R., Crossley N., Whitehouse C.J., Robinson M., Reynolds A.J., Demarchez M., Porter R.M., Shapiro L., and Christiano A.M. 2004. The *lanceolate hair* rat phenotype results from a missense mutation in a calcium coordinating site of the *desmoglein 4* gene. *Genomics* **83:** 747–756.

Kaiser J. 2005. Genomics: Celera to end subscriptions and give data to public GenBank. *Science* **308:** 775.

Karolchik D., Baertsch R., Diekhans M., Furey T.S., Hinrichs A., Lu Y.T., Roskin K.M., Schwartz M., Sugnet C.W., Thomas D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.

Korstanje R. and DiPetrillo K. 2004. Unraveling the genetics of chronic kidney disease using animal models. *Am. J. Physiol. Renal Physiol.* **287:** F347–F352.

Korstanje R. and Paigen B. 2002. From QTL to gene: The harvest begins. *Nat. Genet.* **31:** 235–236.

Kotchen T.A., Zhang H.Y., Covelli M., and Blehschmidt N. 1991. Insulin resistance and blood pressure in Dahl rats and in one-kidney, one-clip hypertensive rats. *Am. J. Physiol.* **261:** E692–E697.

Kuramoto T., Kuwamura M., and Serikawa T. 2004. Rat neurological mutations *cerebellar vermis defect* and *hobble* are caused by mutations in the netrin-1 receptor gene *Unc5h3*. *Brain Res. Mol. Brain Res.* **122:** 103–108.

Kwitek A.E., Jacob H.J., Baker J.E., Dwinell M.R., Forster H.V., Greene A.S., Kunert M.P., Lombard J.H., Mattson D.L., Pritchard K.A., Jr., et al. 2006. BN phenome: Detailed characterization of the cardiovascular, renal, and pulmonary systems of the sequenced rat. *Physiol. Genomics* **25:** 303–313.

Lee M.J., Stephenson D.A., Groves M.J., Sweeney M.G., Davis M.B., An S.F., Houlden H., Salih M.A., Timmerman V., de Jonghe P., et al. 2003. Hereditary sensory neuropathy is caused by a mutation in the delta subunit of the cytosolic chaperonin-containing t-complex peptide-1 (*Cct4*) gene. *Hum. Mol. Genet.* **12:** 1917–1925.

Lee Y., Tsai J., Sunkara S., Karamycheva S., Pertea G., Sultana R., Antonescu V., Chan A., Cheung F., and Quackenbush J. 2005. The TIGR Gene Indices: Clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.* **33:** D71–D74.

Liang M., Yuan B., Rute E., Greene A.S., Olivier M., and Cowley A.W., Jr. 2003. Insights into Dahl salt-sensitive hypertension revealed by temporal patterns of renal medullary gene expression. *Physiol. Genomics* **12:** 229–237.

Liao G., Wang J., Guo J., Allard J., Cheng J., Ng A., Shafer S., Puech A., McPherson J.D., Foernzler D., Peltz G., and Usuka J. 2004. In silico genetics: Identification of a functional element regulating H2-Eα gene expression. *Science* **306:** 690–695.

Liefeldt L., Schonfelder G., Bocker W., Hocher B., Talsness C.E., Rettig R., and Paul M. 1999. Transgenic rats expressing the human ET-2 gene: A model for the study of endothelin actions in vivo. *J. Mol. Med.* **77:** 565–574.

Lindblad-Toh K., Winchester E., Daly M.J., Wang D.G., Hirschhorn J.N., Laviolette J.P., Ardlie K., Reich D.E., Robinson E., Sklar P., et al. 2000. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.* **24:** 381–386.

Luscher T.F., Raij L., and Vanhoutte P.M. 1987. Endothelium-dependent vascular responses in normotensive and hypertensive Dahl rats. *Hypertension* **9:** 157–163.

MacMurray A.J., Moralejo D.H., Kwitek A.E., Rutledge E.A., Van Yserloo B., Gohlke P., Speros S.J., Snyder B., Schaefer J., Bieg S., et al. 2002. Lymphopenia in the BB rat model of type 1 diabetes is due to a mutation in a novel immune-associated nucleotide (*Ian*)-related gene. *Genome Res.* **12:** 1029–1039.

Malek R.L., Wang H.Y., Kwitek A.E., Greene A.S., Bhagabati N., Borchardt G., Cahill L., Currier T., Frank B., Fu X., et al. 2006. Physiogenomic resources for rat models of heart, lung and blood disorders. *Nat. Genet.* **38:** 234–239.

Mashimo T., Birger V., Kuramoto T., and Serikawa T. 2005. Rat Phenome Project: The untapped potential of existing rat strains. *J. Appl. Physiol.* **98:** 371–379.

Monti J., Gross V., Luft F.C., Franca Milia A., Schulz H., Dietz R., Sharma A.M., and Hubner N. 2001. Expression analysis

using oligonucleotide microarrays in mice lacking bradykinin type 2 receptors. *Hypertension* **38:** E1–E3.

Moreno C., Dumas P., Kaldunski M.L., Tonellato P.J., Greene A.S., Roman R.J., Cheng Q., Wang Z., Jacob H.J., and Cowley A.W., Jr. 2003. Genomic map of cardiovascular phenotypes of hypertension in female Dahl S rats. *Physiol. Genomics* **15:** 243–257.

Mott R., Talbot C.J., Turri M.G., Collins A.C., and Flint J. 2000. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci.* **97:** 12649–12654.

Mullins J.J., Peters J., and Ganten D. 1990. Fulminant hypertension in transgenic rats harbouring the mouse Ren-2 gene. *Nature* **344:** 541–544.

Murphy J.M., Stewart R.B., Bell R.L., Badia-Elder N.E., Carr L.G., McBride W.J., Lumeng L., and Li T.K. 2002. Phenotypic and genotypic characterization of the Indiana University rat lines selectively bred for high and low alcohol preference. *Behav. Genet.* **32:** 363–388.

Nadeau J.H. 2000. Muta-genetics or muta-genomics: The feasibility of large-scale mutagenesis and phenotyping programs. *Mamm. Genome* **11:** 603–607.

Nolan P.M., Peters J., Strivens M., Rogers D., Hagan J., Spurr N., Gray I.C., Vizor L., Brooker D., Whitehill E., et al. 2000. A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat. Genet.* **25:** 440–443.

Oiso N., Riddle S.R., Serikawa T., Kuramoto T., and Spritz R.A. 2004. The rat Ruby (*R*) locus is *Rab38*: Identical mutations in Fawn-hooded and Tester-Moriyama rats derived from an ancestral Long Evans rat sub-strain. *Mamm. Genome* **15:** 307–314.

Okimoto K., Sakurai J., Kobayashi T., Mitani H., Hirayama Y., Nickerson M.L., Warren M.B., Zbar B., Schmidt L.S., and Hino O. 2004. A germ-line insertion in the Birt-Hogg-Dubé (*BHD*) gene gives rise to the Nihon rat model of inherited renal cancer. *Proc. Natl. Acad. Sci.* **101:** 2023–2027.

Olofsson P., Holmberg J., Pettersson U., and Holmdahl R. 2003a. Identification and isolation of dominant susceptibility loci for pristane-induced arthritis. *J. Immunol.* **171:** 407–416.

Olofsson P., Holmberg J., Tordsson J., Lu S., Akerstrom B., and Holmdahl R. 2003b. Positional identification of *Ncf1* as a gene that regulates arthritis severity in rats. *Nat. Genet.* **33:** 25–32.

Ovcharenko I., Nobrega M.A., Loots G.G., and Stubbs L. 2004. ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res.* **32:** W280–W286.

Pandey J., Cracchiolo D., Hansen F.M., and Wendell D.L. 2002. Strain differences and inheritance of angiogenic versus angiostatic activity in oestrogen-induced rat pituitary tumours. *Angiogenesis* **5:** 53–66.

Petersen G., Johnson P., Andersson L., Klinga-Levan K., Gómez-Fabre P.M., and Ståhl F. 2005. RatMap—Rat genome tools and data. *Nucleic Acids Res.* **33:** D492–D494.

Pinto-Sietsma S.J. and Paul M. 1997. Transgenic rats as models for hypertension. *J. Hum. Hypertens.* **11:** 577–581.

Pletcher M.T., McClurg P., Batalov S., Su A.I., Barnes S.W., Lagler E., Korstanje R., Wang X., Nusskern D., Bogue M.A., et al. 2004. Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol.* **2:** e393.

Pravenec M., Klir P., Kren V., Zicha J., and Kunes J. 1989. An analysis of spontaneous hypertension in spontaneously hypertensive rats by means of new recombinant inbred strains. *J. Hypertens.* **7:** 217–221.

Pravenec M., Kren V., Krenova D., Bila V., Zidek V., Simakova M., Musilova A., van Lith H.A., and van Zutphen L.F. 1999. HXB/Ipcv and BXH/Cub recombinant inbred strains of the rat: Strain distribution patterns of 632 alleles. *Folia Biol.* **45:** 203–215.

Pravenec M., Gauguier D., Schott J.J., Buard J., Kren V., Bila V., Szpirer C., Szpirer J., Wang J.M., Huang H., et al. 1995. Mapping of quantitative trait loci for blood pressure and cardiac mass in the rat by genome scanning of recombinant inbred strains. *J. Clin. Invest.* **96:** 1973–1978.

Pravenec M., Gauguier D., Schott J.J., Buard J., Kren V., Bila V., Szpirer C., Szpirer J., Wang J.M., Huang H., et al. 1996. A genetic linkage map of the rat derived from recombinant inbred strains. *Mamm. Genome* **7:** 117–127.

Pravenec M., Landa V., Zidek V., Musilova A., Kren V., Kazdova L., Aitman T.J., Glazier A.M., Ibrahimi A., Abumrad N.A., et al. 2001. Transgenic rescue of defective Cd36 ameliorates insulin resistance in spontaneously hypertensive rats. *Nat. Genet.* **27:** 156–158.

Pravenec M., Zidek V., Musilova A., Simakova M., Kostka V., Mlejnek P., Kren V., Krenova D., Bila V., Mikova B., et al. 2002. Genetic analysis of metabolic defects in the spontaneously hypertensive rat. *Mamm. Genome* **13:** 253–258.

Provoost A.P. 1994. Spontaneous glomerulosclerosis: Insights from the fawn-hooded rat. *Kidney Int. Suppl.* **45:** S2–S5.

Rapp J.P. 1982. Dahl salt-susceptible and salt-resistant rats. A review. *Hypertension* **4:** 753–763.

———. 2000. Genetic analysis of inherited hypertension in the rat. *Physiol. Rev.* **80:** 135–172.

Reaven G.M., Twersky J., and Chang H. 1991. Abnormalities of carbohydrate and lipid metabolism in Dahl rats. *Hypertension* **18:** 630–635.

Robinson R. 1987. Genetic linkage in the Norway rat. *Genetica* **74:** 137–142.

Roman R.J. and Kaldunski M. 1991. Pressure natriuresis and cortical and papillary blood flow in inbred Dahl rats. *Am. J. Physiol.* **261:** R595–R602.

Rost S., Fregin A., Ivaskevicius V., Conzelmann E., Hortnagel K., Pelz H.J., Lappegard K., Seifried E., Scharrer I., Tuddenham E.G., et al. 2004. Mutations in *VKORC1* cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature* **427:** 537–541.

Roth M.P., Viratelle C., Dolbois L., Delverdier M., Borot N., Pelletier L., Druet P., Clanet M., and Coppin H. 1999. A genome-wide search identifies two susceptibility loci for experimental autoimmune encephalomyelitis on rat chromosomes 4 and 10. *J. Immunol.* **162:** 1917–1922.

Rubattu S., Volpe M., Kreutz R., Ganten U., Ganten D., and Lindpaintner K. 1996. Chromosomal mapping of quantitative trait loci contributing to stroke in a rat model of complex human disease. *Nat. Genet.* **13:** 429–434.

Scheetz T.E., Laffin J.J., Berger B., Holte S., Baumes S.A., Brown R., II, Chang S., Coco J., Conklin J., Crouch K., et al. 2004. High-throughput gene discovery in the rat. *Genome Res.* **14:** 733–741.

Sebastiani P., Lazarus R., Weiss S.T., Kunkel L.M., Kohane I.S., and Ramoni M.F. 2003. Minimal haplotype tagging. *Proc. Natl. Acad. Sci.* **100:** 9900–9905.

Segre J.A., Nemhauser J.L., Taylor B.A., Nadeau J.H., and Lander E.S. 1995. Positional cloning of the nude locus: Genetic, physical, and transcription maps of the region and mutations in the mouse and rat. *Genomics* **28:** 549–559.

Shisa H., Lu L., Katoh H., Kawarai A., Tanuma J., Matsushima Y., and Hiai H. 1997. The LEXF: A new set of rat recombinant inbred strains between LE/Stm and F344. *Mamm. Geno*
324–327.

Smits B.M., Mudde J., Plasterk R.H., and Cuppen E. 200
selected mutagenesis of the rat. *Genomics* **83:** 33

Smits B.M.G., van Zutphen B.F.M., Plasterk R.H.A.
2004b. Genetic variation in coding regi
within commonly used inbred rat strain
1285–1290.

Smits B.M.G., Guryev V., Zeegers D., Wedek
and Cuppen E. 2005. Efficient single nuc
discovery in laboratory rat strains usir
candidates. *BMC Genomics* **6:** 170.

Smits B.M.G., Mudde J.B., van de Belt J
Homberg J., Guryev V., Cools A.R.,
R.H.A., and Cuppen E. 2006. Gene
and mutant models in the labor

Target-selected mutagenesis. *Pharmacogenet. Genomics* **16:** 159–169.

Snell G. 1948. Methods for the study of histocompatibility genes. *J. Genet.* **49:** 87–108.

Steen R.G., Kwitek-Black A.E., Glenn C., Gullings-Handley J., Van Etten W., Atkinson O.S., Appel D., Twigger S., Muir M., Mull T., et al. 1999. A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Res.* **9:** AP1–8, insert.

Stephens M., Sloan J.S., Robertson P.D., Scheet P., and Nickerson D.A. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.* **38:** 375–381.

Stoll M., Cowley A.W., Jr., Tonellato P.J., Greene A.S., Kaldunski M.L., Roman R.J., Dumas P., Schork N.J., Wang Z., and Jacob H.J. 2001. A genomic-systems biology map for cardiovascular function. *Science* **294:** 1723–1726.

Stoll M., Kwitek-Black A.E., Cowley A.W., Jr., Harris E.L., Harrap S.B., Krieger J.E., Printz M.P., Provoost A.P., Sassard J., and Jacob H.J. 2000. New target regions for human hypertension via comparative genomics. *Genome Res.* **10:** 473–482.

Stylianou I.M., Christians J.K., Keightley P.D., Bünger L., Clinton M., Bulfield G., and Horvat S. 2004. Genetic complexity of an obesity QTL (*Fob3*) revealed by detailed genetic mapping. *Mamm. Genome* **15:** 472–481.

Sugiyama F., Churchill G.A., Higgins D.C., Johns C., Makaritsis K.P., Gavras H., and Paigen B. 2001. Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics* **71:** 70–77.

Svenson K.L., Cheah Y.C., Shultz K.L., Mu J.L., Paigen B., and Beamer W.G. 1995. Strain distribution pattern for SSLP markers in the SWXJ recombinant inbred strain set: Chromosomes 1 to 6. *Mamm. Genome* **6:** 867–872.

Tanomura H., Miyake T., Taniguchi Y., Manabe N., Kose H., Matsumoto K., Yamada T., and Sasaki Y. 2002. Detection of a quantitative trait locus for intramuscular fat accumulation using the OLETF rat. *J. Vet. Med. Sci.* **64:** 45–50.

Tian X.L., Pinto Y.M., Costerousse O., Franz W.M., Lippoldt A., Hoffmann S., Unger T., and Paul M. 2004. Over-expression of angiotensin converting enzyme-1 augments cardiac hypertrophy in transgenic rats. *Hum. Mol. Genet.* **13:** 1441–1450.

Toga A.W., Santori E.M., Hazani R., and Ambach K. 1995. A 3D digital map of rat brain. *Brain Res. Bull.* **38:** 77–85.

Twigger S.N., Pasko D., Nie J., Shimoyama M., Bromberg S., Campbell D., Chen J., Dela Cruz N., Fan C., Foote C., et al. 2005. Tools and strategies for physiological genomics: The Rat Genome Database. *Physiol. Genomics* **23:** 246–256.

Visscher P.M. 1999. Speed congenics: Accelerated genome recovery using genetic markers. *Genet. Res.* **74:** 81–85.

Vitt U., Gietzen D., Stevens K., Wingrove J., Becha S., Bulloch S., Burrill J., Chawla N., Chien J., Crawford M., et al. 2004. Identification of candidate disease genes by EST alignments, synteny, and expression and verification of Ensembl genes on rat chromosome 1q43-54. *Genome Res.* **14:** 640–650.

Wade C.M., Kulbokas E.J., III, Kirby A.W., Zody M.C., Mullikin [illegible]C., Lander E.S., Lindblad-Toh K., and Daly M.J. 2002. The [illegible]saic structure of variation in the laboratory mouse [illegible]me. *Nature* **420:** 574–578.

[illegible]orstanje R., Higgins D., and Paigen B. 2004. [illegible] analysis in multiple crosses to identify a QTL [illegible]*e Res.* **14:** 1767–1772.

Ward C.J., Hogan M.C., Rossetti S., Walker D., Sneddon T., Wang X., Kubly V., Cunningham J.M., Bacallao R., Ishibashi M., et al. 2002. The gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein. *Nat. Genet.* **30:** 259–269.

Ways J.A., Cicila G.T., Garrett M.R., and Koch L.G. 2002. A genome scan for loci associated with aerobic running capacity in rats. *Genomics* **80:** 13–20.

Weber J.L. and P.E. May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44:** 388–396.

Wendell D.L. and Gorski J. 1997. Quantitative trait loci for estrogen-dependent pituitary tumor growth in the rat. *Mamm. Genome* **8:** 823–829.

Wilder S.P., Bihoreau M.-T., Argoud K., Watanabe T.K., Lathrop M., and Gauguier D. 2004. Integration of the rat recombination and EST maps in the rat genomic sequence and comparative mapping analysis with the mouse genome. *Genome Res.* **14:** 758–765.

Wiltshire T., Pletcher M.T., Batalov S., Barnes S.W., Tarantino L.M., Cooke M.P., Wu H., Smylie K., Santrosyan A., Copeland N.G., et al. 2003. Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc. Natl. Acad. Sci.* **100:** 3380–3385.

Yalcin B., Fullerton J., Miller S., Keays D.A., Brady S., Bhomra A., Jefferson A., Volpi E., Copley R.R., Flint J., and Mott R. 2004a. Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc. Natl. Acad. Sci.* **101:** 9734–9739.

Yalcin B., Willis-Owen S.A., Fullerton J., Meesaq A., Deacon R.M., Rawlins J.N., Copley R.R., Morris A.P., Flint J., and Mott R. 2004b. Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. *Nat. Genet.* **36:** 1197–1202.

Yeung R.S., Xiao G.H., Jin F., Lee W.C., Testa J.R., and Knudson A.G. 1994. Predisposition to renal carcinoma in the Eker rat is determined by germ-line mutation of the tuberous sclerosis 2 (*TSC2*) gene. *Proc. Natl. Acad. Sci.* **91:** 11413–11416.

Yokoi N., Namae M., Wang H.Y., Kojima K., Fuse M., Yasuda K., Serikawa T., Seino S., and Komeda K. 2003. Rat neurological disease *creeping* is caused by a mutation in the *reelin* gene. *Brain Res. Mol. Brain Res.* **112:** 1–7.

Yokoi N., Komeda K., Wang H.Y., Yano H., Kitada K., Saitoh Y., Seino Y., Yasuda K., Serikawa T., and Seino S. 2002. *Cblb* is a major susceptibility gene for rat type 1 diabetes mellitus. *Nat. Genet.* **31:** 391–394.

Zan Y., Haag J.D., Chen K.S., Shepel L.A., Wigington D., Wang Y.R., Hu R., Lopez-Guajardo C.C., Brose H.L., Porter K.I., et al. 2003. Production of knockout rats using ENU mutagenesis and a yeast-based screening assay. *Nat. Biotechnol.* **21:** 645–651.

Zhou Q., Renard J.P., Le Friec G., Brochard V., Beaujean N., Cherifi Y., Fraichard A., and Cozzi J. 2003. Generation of fertile cloned rats by regulating oocyte activation. *Science* **302:** 1179.

Zidek V., Pintir J., Musilova A., Bila V., Kren V., and Pravenec M. 1999. Mapping of quantitative trait loci for seminal vesicle mass and litter size to rat chromosome 8. *J. Reprod. Fertil.* **116:** 329–333.

Zimdahl H., Nyakatura G., Brandt P., Schulz H., Hummel O., Fartmann B., Brett D., Droege M., Monti J., Lee Y.A., et al. 2004. A SNP map of the rat genome generated from cDNA sequences. *Science* **303:** 807.

# 30 猫

Monika J. Lipinski, Nicholas Billings, and Leslie A. Lyons

*Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, California 95616*



## 简介

家猫（*Felis catus*）的基因组种存在着很多秘密仍没有被发现；在其他物种中，这些同样神秘之处已经被揭示了出来。家猫的起源是一个谜，很多家猫如此漂亮的原因也有待研究。家猫基因组学从技术先进的人类基因组计划中受益匪浅，现在，在基因组测序问题上，猫是第二种被人们关注的食肉动物物种。最近，家猫的一个低覆盖范围、2X序列已经被成功获得，但是进一步更深入的研究和在其他品种和个体中的重新定序研究仍处在起步水平。如果不考虑序列范围问题的话，那么猫遗传学和基因组学已经得到并且在不断快速发展着。很多猫类疾病和表型突变已经在过去的几年里被陆续鉴定了出来。猫类独特的品种发展过程和基因组组成上的差异都表明猫类不仅仅是一种小型的狗而已。在这一章，我们探讨了猫科动物发展和动态的细微差别，并呈现了可以成为猫类特异研究关注点的具有一般性和特有性的生物学、生理学方面。这些研究无疑在将来可以促进猫类、人类和其他物种的健康。

## 猫的起源

家猫是36个现存的猫科动物的一种（Kitchener 1991；Seidensticker and Lumpkin 1991；Sunquist and Sunquist 2002），它和其他4个古代世界的野猫种类一起组成家猫“猫属”的世系。这个世系包括中国山猫（*F. bieti*）、沙猫（*F. margarita*）、黑足猫

(*F. nigripes*)、丛林猫（*F. chaus*）(Johnson et al. 2006)。另外还有两种野猫也经常被认为是不同的物种并且和家猫有明显的区别，这两种野猫包括欧洲野猫(*F. silvestis*) 和非洲野猫（*F. libyca*)。然而，这两种野猫和家猫在杂交后可以产生可育杂种，而且这些小型野猫和家猫的物种形成问题仍然是一个谜，特别是因为遗传学研究还不能解释他们在近期内所产生的那些分歧的原因（图 30-1)。很多其他欧洲野猫、亚洲野猫和非洲野猫的亚种都已经被描述出来。然而，真正的起源和导致猫类驯化的事件仍没有被解决，是一个谜。

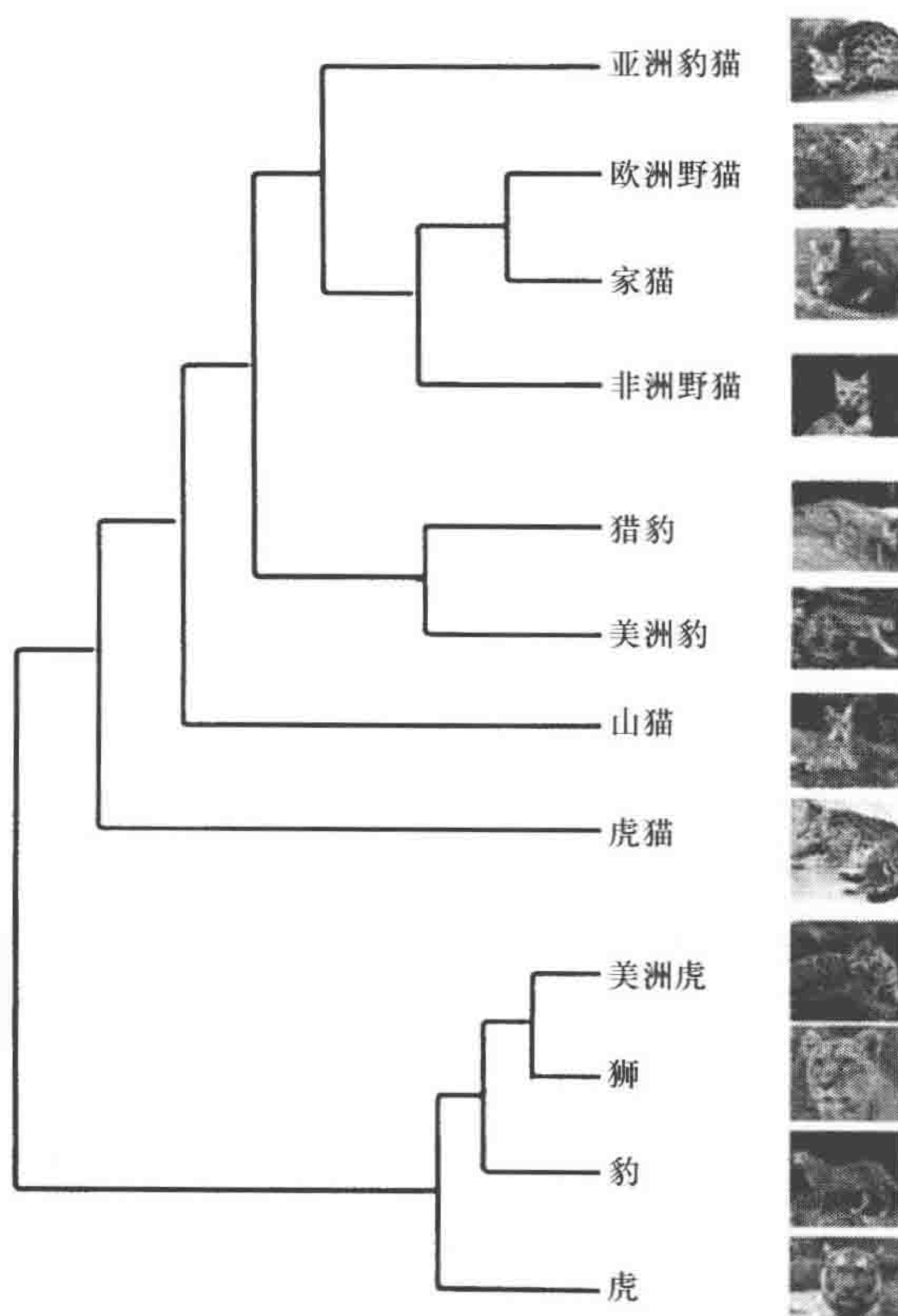



图 30-1　小型野猫的系统发生关系（Johnson 2006)

当农业发展地点和粮谷储藏的驯化过程已经越发清楚的时候，猫类驯化问题开始成为焦点问题。人们已经了解到了以前的一些独立的农业发展的区域，包括新月沃土区域，尤其是公元前 10 000 年和公元前 9000 年的叙利亚地区。大概在新月沃土最早的农业定居者出现的 5000 年后，农业在埃及的尼罗河谷也开始出现。同时，大量的作物已经在更早的几千年前被引入亚洲和印度次大陆上（Zohary and Hopf 2002)，这暗示着埃及也许是猫类驯化比较早的区域之一。埃及人记录了大量关于他们文化的信息，在他们的社会中不乏“家猫”出现的情况。所以，很多人认为埃及人是最早驯化猫类的。最早对于猫类的艺术表现出现在公元前 2500 年左右，但是显然，制造猫的木乃伊的过程直到托勒密时期的晚期才出现，大概在公元前 500 年到公元 100 年左右

(Malek 2006)。

几种小型的野猫物种，即非洲野猫、亚洲野猫、欧洲野猫，在所有早期的农业地域中都已经存在，并且很可能就是家猫的祖先。随着人类进行了从捕猎、收集食物到更稳定的定居生活模式的转变，永久的定居现象也开始慢慢涌现。村庄产生的成堆的垃圾和储存的粮食吸引来了各种鼠类，而这些鼠类都是小型野猫的猎物。所以，很有可能猫类实际上积极地参与到了人们对它们的驯化过程中，人类和猫科动物发展了一种共生和相互的宽容关系。接受了人类存在的猫类得到了更稳定的食物来源。人类接受猫科动物的事实也帮助人类控制了瘟疫的蔓延和减少了谷类储藏的污染情况，从而控制了通过啮齿类动物传播的动物传染病。但是，又是从什么时候开始猫类开始寻求人类的感情和陪伴？什么时候人类制订了第一个控制猫类繁殖的计划？这些问题的答案仍然有待探究。不管猫类育种是在什么地方或者什么时候开始发展的，猫都是在历史的近期才被驯化成为我们的陪伴动物，这与其他在更早时期便被驯化的陪伴动物不同，如家养的狗。所以，在猫类育种过程中遗传改变的动态变化很可能与其他具有更古老驯化历史和更长育种历史的物种明显不同。

## 猫类品种

猫类育种的动态过程和人类的其他陪伴动物和农业物种有明显不同。这些差别对于适当的遗传工具、资源和技术的发展非常重要，这些工具、资源和技术都会成为猫类遗传研究最有效和最有价值的工具和手段。在全球范围，可以代表绝大多数猫类的品系是随机交配品系和野生猫类，而不是那些观赏品种。概观世界上的猫类分布，美国占有最高比例的纯种猫类品种。加州大学戴维斯分校兽医教学医院每年接待超过 20 000 个猫类病患；然后，其中只有 10％～15％属于纯种猫类（Louwerens et al. 2005）。所以，对于复杂性状的种群研究可能需要在许多随机交配品系种群中设定标记物和进行连锁不平衡评估。

首个被记载的体现了猫的美学价值的展览是 1871 年在伦敦水晶宫举办的。那次的首届竞赛中只有少数品种参加，包括波斯猫、阿比西尼亚猫和暹罗猫。在美国，猫类爱好者协会（CFA）登记在案的第一只猫是在 1905 年登记的，并且该缅因猫为一个额外的美国品种的猫。很多大百科全书中在家猫的列表里罗列了全球 50～80 个猫的品种（Fogle 1997；Morris 1999）。然而，大部分的品种都是在过去的 50 年间产生的。很多被列出的品种还都没有形成可育种群，所以，也就因为没有子孙而没有存留下来。最重要的猫类品种都在表 30-1 中列出。全球大多数猫类爱好者协会可识别和承认的猫类品种有 35～40 个。然而，只有一部分品种占有品种种群的绝大部分比重。波斯猫及其相关的品种，如一种名为 Exotics 的短毛波斯猫变种，是世界范围内猫类品种中最常见的品种，它们占了纯种猫数量的大部分。尽管可能只有 20％～30％被培育出来的猫进行了注册，但是在 CFA（世界上最大的猫类登记处之一）每年都有大概 40 000 只纯种猫新登记在案。其中有 16 000～20 000 只是波斯猫，最多有 3000 只是外来猫。所以，波斯猫的群体接近了猫类爱好者所养猫总数的 50％。常见的品种中，每年至少有 1000

只注册的阿比尼西亚猫、缅因猫和暹罗猫。其他的常见品系还包括缅甸巴曼猫和缅甸猫。这些流行品种的大部分常常也代表了全世界最古老和最稳定存在的猫类品种。所以，遗传工具和 SNP 研究应该首先着眼于家猫和一些观赏猫品种。

另外，基础品种可能必须要利用基因应用方法进行评估。很多品种都是从一些更古老的品种中衍生而来的，从而形成了一个品种的家系。至多有 17 个品种可以被认为是“基础”或者“天然”品种（表 30-1），这就暗示了很多其他品系都是从这些基础猫类中演化而来。这些演化而来的品种常常是单一基因改变的结果，如长毛和短毛的不同，或者甚至出现无毛的变异，就像在德文王猫（Devon rex）和斯芬克斯（Sphynx）群体中发现的情况一样。颜色的不同也趋向于用来区分品种，如波斯猫的“斑点”变种，这种猫被猫类的狂热爱好者称为喜马拉雅猫并被一些组织认为是一个独立的品种。很多猫类品种来自于单一基因的性状［如苏格兰折耳猫（Scottish fold）的折叠状耳朵和美国反耳猫（American curl）的驼背］，并且在后来发展成了一个在结构上更特殊的品种。最近被鉴定出来的自发突变通常是在随机交配的猫类种群中被发现的，并且通常伴随有

**表 30-1　传统猫类品种和品系家族**

| 品种 | 起源 | Est'd 日期 | 衍生品种 | 头型 |
|---|---|---|---|---|
| Abyssinian[a] | India | 1868 | Somali[b] | mesaticephalic |
| American bobtail | mutation–U.S. | 1960 | | mesaticephalic |
| American curl | mutation–U.S. | 1981 | | mesaticephalic |
| American shorthair | U.S. | 1966 | | brachycephalic |
| American wirehair | mutation–U.S. | 1966 | | mesaticephalic |
| Australian mist | mix–Australia | 1990s | | mesaticephalic |
| Birman[a] | Burma | <1868 | Snowshoe[b] | mesaticephalic |
| British shorthair[a] | England | 1870s | | brachycephalic |
| Burmese[a] | Burma | 1350–1767 | (Asian) Bombay, Tiffanie,[b] Malayan, Burmilla | brachycephalic |
| Cornish rex | mutation–U.K. | 1950 | | dolichocephalic |
| Chartreux[a] | France | 1300 | | mesaticephalic |
| Devon rex | mutation–U.K. | 1960 | Sphynx (1966) | mesaticephalic |
| Egyptian Mau[a] | Egypt | 1953 | | mesaticephalic |
| European shorthair | Europe | | | brachycephalic |
| Japanese bobtail[a] | Japan | 500–1100 | | mesaticephalic |
| Korat[a] | Thailand | 1350–1767 | | mesaticephalic |
| LaPerm | mutation–U.S. | 1986 | | mesaticephalic |
| Maine coon[a] | U.S. | 1860s | | mesaticephalic |
| Manx | Isle of Man | <1868 | Cymric[b] | mesaticephalic |
| Munchkin | U.S. | 1990s | | mesaticephalic |
| Norwegian forest[a] | Norway | <1868 | | mesaticephalic |
| Ocicat | crossbred | 1964 | Siamese x Abyssinian | mesaticephalic |
| Ojos Azules | mutation | 1980s | | mesaticephalic |
| Persian[a] | Persia | <1868 | Exotic,[b] Kashmir, Himalayan, Peke-faced, Burmilla | brachycephalic |
| Russian blue[a] | Russia | <1868 | Nebelung[b] | mesaticephalic |
| Ragdoll | selection | 1960s | Ragamuffin | |
| Scottish Fold | mutation | 1961 | Highland fold[b] (Coupari) | brachycephalic |
| Selkirk rex | mutation–U.S. | 1980s | | mesaticephalic |
| Siamese[a] | Thailand | 1350–1767 | Colorpoint,[b] Javanese,[b] Balinese,[b] Oriental,[b] Havana brown, Don Sphynx | dolichocephalic |
| Siberian[a] | Russia | <1868 | | mesaticephalic |
| Sokoke[a] | Africa | | | mesaticephalic |
| Tonkinese | crossbred | 1950s | Siamese x Burmese | brachycephalic |
| Turkish Angora[a] | Ankara | 1400 | | mesaticephalic |
| Turkish Van[a] | Van Lake | <1868 | | mesaticephalic |

注：A 表示基础品种或天然品种；B 表示很多衍生品种都是基础品种的长毛或者短毛变异体，但是都有不同的品种名称；其他的都在品种名称中标明长毛或者短毛。至少有 10 个额外的 rex 卷毛性状毛皮的种群还没有发展成可育种群或者出现明显特征。

各种形态样式所需的品种组合。所以，很多新品种和一些古老的品种都存在可允许的异种繁殖来影响其“类型”并帮助产生该品种基础上的遗传多样性。波斯猫有一个人们很欣赏的短头类型，所以，这种倾向性也影响了其他的很多品种。喜欢长头类型的培育者通常会用暹罗猫家族进行异种繁殖。任何品种所产生的异种繁殖品种在猫类登记处之间都会有所不同，而且同一个品种由于国家的不同也会有不同的名字。例如，在英国猫迷管理委员会（GCCF）和欧洲的猫类国际联合会（FIFe）被命名为缅甸猫的品种在美国被人们称为外来缅甸猫品种，而且这些猫类“品种”在不同国家也具有明显不同的颅面类型。哈瓦那雪茄色棕猫在美国发展出了一个很有特色的品种，并且相对于其原始品种、暹罗猫和东方短毛猫来看也具有一个明显不同的颅面结构。然而，在欧洲，东方短毛猫的栗色变种和哈瓦那雪茄色棕猫很类似。有些品种，如克拉特猫（Korats）和土耳其梵猫（Turkish vans），在全世界所有的注册点都有相似的判断标准。奇怪的是，有些猫类品种实际上是一些明显不同的猫类品种和家猫的杂交种。亚洲美洲豹纹猫是现在非常流行的孟加拉猫的祖先之一，但是它却没有被 CFA 注册。有些被称为热带草原猫的杂交种和有些被称为 chaussies 的丛林猫杂交种现在也处在越来越流行的趋势当中。所以，基因组工具应当被更多的用于研究这三种猫类物种，从而通过对这些猫类杂交种的研究来支持疾病的研究。

所以，从系统发生学角度来看，进行猫类品种培育的人更像是“分配者”（splitter）而不是单纯的“装卸者”（lumper），而且他们只利用了一部分基础品种就完成了大多数猫类品种的改变。在一个对 19 个猫类品种进行的 39 个微卫星标记物的分析评估中，人们发现 12 个随机交配品系和 3 个野生猫类亚种在其祖先品系中存在一些基本的关系（图 30-2），在图中用 Phylip 软件将这种关系用邻接法（N-J）绘制成发生树进行表示。Cavalli-Sforza 的弦测量和 Nei 的遗传距离所获得的结果是相似的，但是，因为通过弦距离（cord distance）获得的 N-J 树有更高的自引导值（图 30-2），所以该方法是对上述存在的种关系更有效的证明。[自举法（bootstrap）是一种用来评估一个系统发生树可信度的重取样（关于数据的假重复问题）方法。当数据被重新取样和系统树被构建的时候，对应于某个特定分支模式上的百分数数值就会出现]。所显示出来的这些关系表明，那些被记载为演化物种的品种（如新加坡猫、缅甸猫或者哈瓦那雪茄色棕猫和暹罗猫）在发生树上都是非常紧密地簇集在一起的。东非新培育出来的品种，如肯尼亚猫（Sokoke）都和其野生的祖先非常相似，他们的野生祖先是来自于肯尼亚群岛拉姆岛和佩特岛的猫类。另外，有三类猫在发生树上是明显聚集在一起的：来自于远东的猫类，如野猫和肯尼亚棕猫、新加坡猫、呵叻猫和缅甸猫；受阿拉伯地区影响的猫类，如野生猫类和肯尼亚猫；来自地中海的猫类，这类猫的品种几乎包括了所有其他的品种和种群。

通过 Structure 软件进行了 19 个品种的贝叶斯聚集分析，科学家获得了猫类品种的其他结构（Pitchard et al. 2000）。假设这 19 个物种都是在遗传角度上不同的种群，被推断出的种群数量设置为 10（K=19）。这种方法可以将猫类分成几个自动推断出的簇，这些簇对应于它们各自种群的特征，在分配成簇后都具有很高的后验概率，并且这些自动推断出来的簇可以正确地将个体分配到它们特定的品种中，正确率高达 98%（图 30-

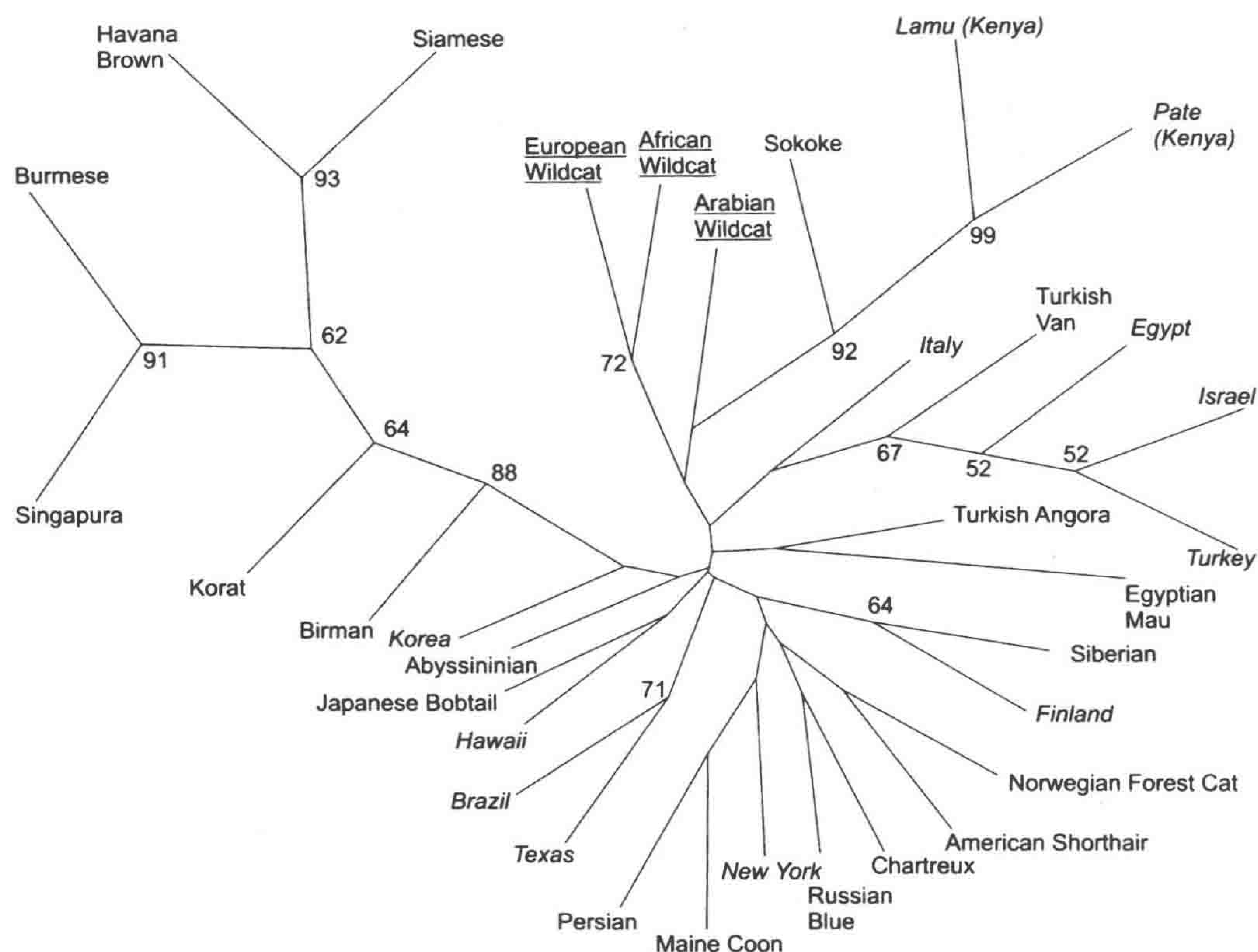



图 30-2　猫类品种系统发生分析。邻接法获得的发生树利用 Phylip 软件构建（Felsenstein 1989）。图中表示利用弦距离表示的树状图。连接点处的数字表示最高自引导值。随机交配品系种群用斜体表示，野猫种群用划线表示

3）。大多数被分析的品种都具有很强的群集性，来源于相似地区的品种也共享着很多相同的等位基因。远东的品种——缅甸猫和新加坡猫，是完全不同的品种，但是确实共享着很多遗传要素。目前仍有 4 个远东品种没有被分析。由于贝叶斯模型不能分离新加坡猫和缅甸猫品种，这也暗示了这两个品种可能源于一个近期的共同起源。一方面，猫类育种的民间传说暗示这两种猫都是缅甸本地的品种。另一方面，比较普遍的观点认为，缅甸猫从美国被带去了新加坡并且与当地猫类品种杂交培育，其后代被带回美国从而成为了新加坡猫品种。此外，哈瓦那雪茄色棕猫和暹罗猫的区分也存在着矛盾性。阿比西尼亚猫表现为一个具有明显区别特点的品种，而更多的欧洲品种在区分上都相对地不明确一些。近几年，猫类的随机交配品系或者野生种群被培育成了更新的、区域特定的品种，如俄国的西伯利亚猫和非洲的。西伯利亚品种和随机交配品系种群相比具有一些遗传变异，就像 Structure 分析所表明的那样（图 30-3）。

## 猫类表型变异

单一基因突变可以产生很多不同的猫类品种，并且很多品种种群就是靠等位基因的

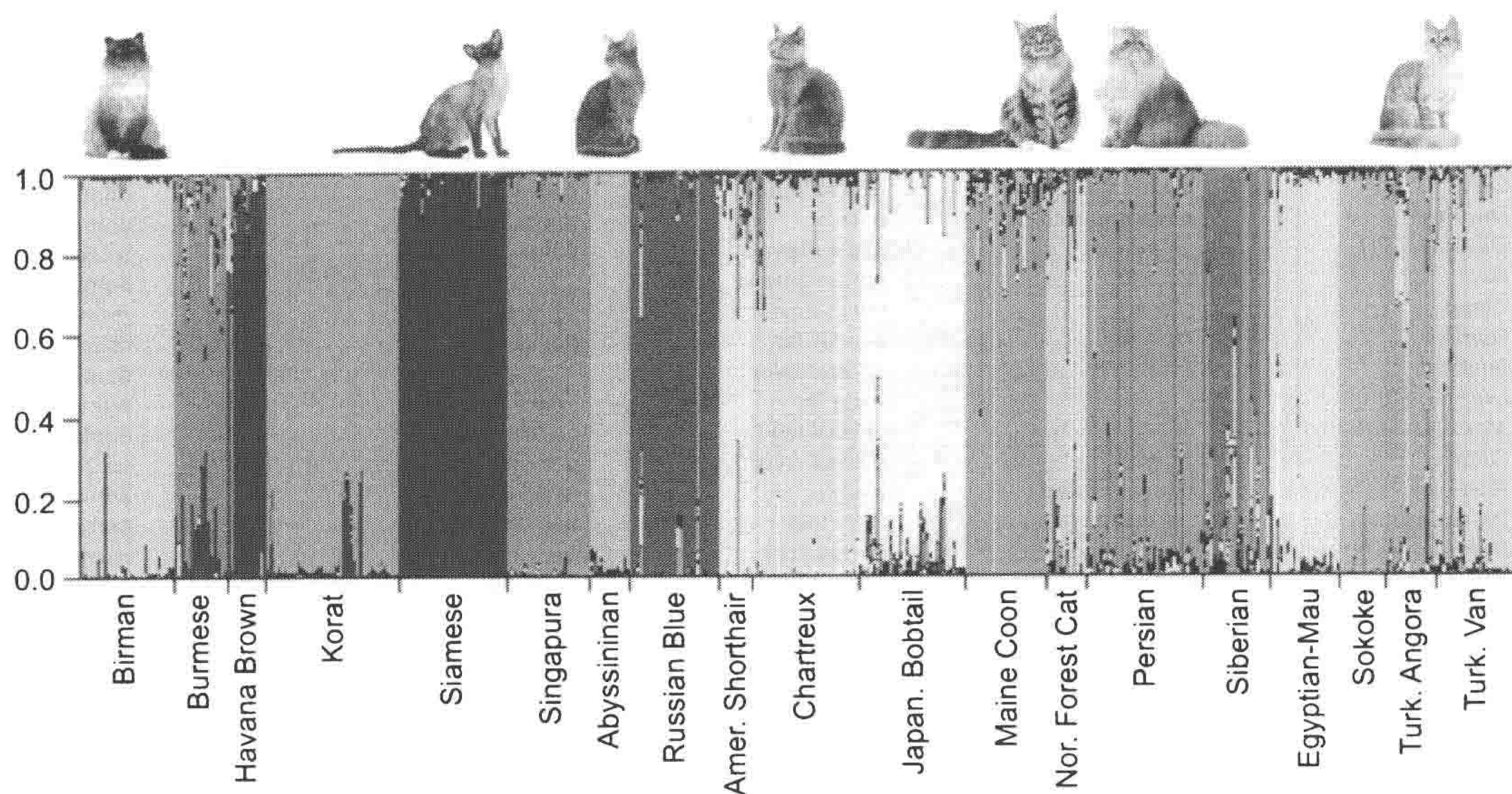



图 30-3 家猫基础来源结构。表示了来自 19 个品种的个体的 39 个微卫星微点的基因型。利用 Structure 软件，品种可以被清楚地区分开来表示这些品种代表了种群的真实数量（K=19）。用相同的颜色表示的猫类品种说明尚不能利用该方法分清，如哈瓦那雪茄色棕猫和暹罗猫或者缅甸猫和新加坡猫

正常和变异来进行区分的。在家猫和相关品种中已经鉴定出有超过 24 个单一基因表型的变异，包括显性的、共显性的、隐性的、性连锁的和纯合子致死等特点的变异（表 30-2）。猫类皮毛颜色的遗传异质性、多能性和上位性效应使得猫类成为了基础水平上和高级水平上遗传的一个绝好的指示标记。尽管在所有物种特定染色体中第一批被绘图的性状就包括猫类性连锁的橙色着色的现象（Ibsen 1916），但是，猫类的首个常染色体上性状连锁的基因却是在 70 年后才被发现。这些被发现的现象有血红素的多态性和由酪氨酸酶突变引起的暹罗猫的图案变化（O' Brien et al. 1986）。现在，用于猫类研究的基因组工具越来越高效，很多猫类表型的遗传机制也正在被解码，其中包括被人称为"点状图案"的暹罗猫图案的遗传机制。这种在暹罗猫和喜马拉雅猫身上发现的点状着色很容易被辨别鉴定，因为这个酪氨酸酶基因会在其他物种中引起类似的表型（Lyons et al. 2005b）。目前，通过候补基因的方法已经鉴别了大多数的猫类表型突变，因为在其他物种中会出现明显的酶促反应机制被破坏现象或者会记录到同样的毛发着色的现象。基于家族连锁的研究也关联到候补基因，如 *KIT* 基因和白色斑点基因（Cooper et al. 2006），但是其他的连锁分析也暗示着还有不存在已知候补基因的遗传区域，如染色体 B1 位点和 *Tabby* 位点（Lyons et al. 2006）。迄今，在不同品种中，所有猫类所出现的色彩突变已经被证明都是一样的，而且这种色彩突变在后代中也会出现同样的特征；而有几个突变可能导致了家猫中出现长毛的现象（Drogemuller et al. 2007）。

**表 30-2　家猫的表型变化**

| 基因座 | 基因 | 表型 | 变异 | 遗传性 |
|---|---|---|---|---|
| *Agouti (ASIP)* | $A > a$ | banded fur | solid, self | recessive |
| *Black (TYRP1)* | $B > b > b^l$ | black pigment | brown pigments | allelic series |
| *Color (TYR)* | $C > c^b > c^s > c$ | normal color | temperature-sensitive colors, albino | allelic series |
| *Curl* | $Cu > cu$ | normal pinnea | curled pinnea | dominant |
| *Dwarfism* | not designated | normal | shortened legs | dominant |
| *Dilute (MLPH)* | $D > d$ | dense pigment | dilute pigment | recessive |
| *Fold* | $Fd > fd$ | normal pinnea | | dominant |
| *Gloves* | $G > g$ | normal color | white feet | recessive |
| *Hairless* | $Hr > hr$ | normal fur | no hair | recessive |
| *Inhibitor* | $I > i$ | normal color | | dominant |
| *Long (FGF5)* | $L > l$ | short fur | long fur | recessive |
| *Manx (tailless)* | $M > m$ | normal tail | no (short) tail | dominant |
| Orange | $X^o, x^o$ | normal color | orangish pigments | sex-linked |
| *Peterbald* | not designated | hairless | normal hair | dominant |
| *Polydactyla* | $Pd > pd$ | extra toes | normal toes | dominant |
| *Rex (Cornish)* | $R > r$ | normal hair | curly hair | recessive |
| *Rex (Devon)* | $R^e > r^e$ | normal hair | curly hair | recessive |
| *Rex (LaPerm)* | not designated | curly hair | normal fur | dominant |
| *Rex (Selkirk)* | not designated | curly hair | normal fur | dominant |
| Rex (wirehair) | not designated | normal hair | curly hair | incomplete |
| *Spotting (KIT?)* | $S, s$ | normal color | ventral white | additive |
| *Tabby* | $T^a > T^m > t^b$ | no pattern | pattern: stripes | allelic series |
| *White* | $W > w$ | all white | normal color | dominant |

注：已知基因突变情况的位点在括弧中标记处基因记号。表型表示显性等位基因，突变型表示隐性等位基因。

## 猫类疾病突变

尽管很多猫类的表型性状的出现看似是因为存在着伴随品系而遗传下来的突变，但是这种通过品种而传承下来的突变被证实应该会更明显一些。在猫类的已有记录中，至少记录了 277 个紊乱现象或者在其他物种中存在着可遗传因素的“表现性状”。其中至少有 46 个被认为在猫类中是单一基因决定的性状。18 个不同的基因引起了人们已知的 24 种突变，这些突变都会导致猫类疾病的产生（表 30-3）。然而，只有 10 种引起疾病的突变在猫类各个品种中是分开的、不同的。有些猫类疾病只在研究所用的群体里存在。尽管有些猫类疾病的突变在某个品种中已经被确认存在，但是并不是所有的疾病对于该品种都有效果或者都有意义，因为这些疾病有可能已经被除去或者是一个不定时发生的突变。

大部分第一批被鉴定出来的猫类疾病突变也是得益于候补基因的方法，因为在其他物种的已知基因和突变中可以找到存在的类似特性。在猫类发现的首个疾病突变是假肥大型肌营养不良（DMD）中肌营养不良蛋白（Winand et al.　1994）的启动子突变和 *beta*-己糖胺酶 *A*（*HEXB*）的 beta 亚基中的一个突变，该突变在人类中发生即会导致溶酶体贮积病（LSD）和山德霍夫症（Muldoon et al.　1994）。在各品种中这两者都不存在普遍的患病情况，所以，也无法构成一个物种的易患病体质。另外一些其他的天生性新陈代谢缺陷或者 LSD 已经在波斯猫和暹罗猫中被发现（表 30-3），并且这些疾病已经在实验种群中被保存下来，但是同样，这些疾病也没有在品种间出现普遍的分离。呵叻猫和波斯猫存在不止一个新陈代谢上的缺陷，MPS VI 种群的猫被发现是 *ARSB* 突变体的复合杂合子（Crawley et al.　1998）。呵叻猫是个拥有很小种群的品种，而且从这

**表 30-3　已知突变的特征和猫的疾病**

| 疾病/颜色 | 基因 | 突变 | 品种 | 参考文献 |
|---|---|---|---|---|
| Agouti | *ASIP* | del122–123 | all breeds | Eizirik 等 (2003) |
| Brown | *TYRP1* | b = C8G<br>$b^l$ = C298T | all breeds | Lyons 等 (2005a) |
| Dilution | *MLPH* | T83del | all breeds | Ishida 等 (2006) |
| Color | *TYR* | $c^b$ = G715T<br>$c^s$ = G940A<br>c = C975del | all breeds | Lyons 等 (2005b);<br>Imes 等 (2006) |
| AB blood type (type B) | *CMAH* | 18indel-53 | all breeds | Bighignoli 等 (2007) |
| Gangliosidosis 1[a] | *GBL1* | G1457C | Korat, Siamese | De Maria 等 (1998) |
| Gangliosidosis 2[a] | *HEXB* | 15 bp del (intron) | Burmese | (unpubl.) |
| Gangliosidosis 2 | *HEXB* | inv1467–1491 | DSH | Martin 等 (2004) |
| Gangliosidosis 2 | *HEXB* | C667T | DSH (Japan) | Kanae 等 (2006) |
| Gangliosidosis 2[a] | *HEXB* | C39del | Korat | Muldoon 等 (1994) |
| Gangliosidosis 2 | *GM2A* | del390–393 | DSH | Martin 等 (2005) |
| Glycogen storage disease IV[a] | *GBE1* | 230 bp ins 5′–6 kb del | Norwegian forest | Fyfe 等 (2007) |
| Hemophilia B | *F9* | G247A | DSH | Goree 等 (2005) |
| Hemophilia B | *F9* | C1014T | DSH | Goree 等 (2005) |
| Hypertrophic cardiomyopathy[a] | *MYBPC* | G93C | Maine coon | Meurs 等 (2005a) |
| Hypertrophic cardiomyopathy | *MYBPC* | C2458T | Ragdolls | Meurs 等 (2007) |
| Lipoprotein lipase deficiency | *LPL* | G1234A | DSH | Ginzinger 等 (1996) |
| Alpha mannosidosis | *LAMAN* | del1748–1751 | Persian | Berg 等 (1997) |
| Mucolipidosis II | *GNPTA* | C2655T | DSH | Giger 等 (2006) |
| Mucopolysaccharidosis I | *IDUA* | del1047–1049 | DSH | He 等 (1999) |
| Mucopolysaccharidosis VI | *ARSB* | T1427C | Siamese | Yogalingam 等 (1996) |
| Mucopolysaccharidosis VI | *ARSB* | G1558A | Siamese | Yogalingam 等 (1998) |
| Mucopolysaccharidosis VII | *GUSB* | A1052G | DSH | Fyfe 等 (1999) |
| Muscular dystrophy | *DMD* | 900 bp del M promoter–exon 1 | DSH | Winand 等 (1994) |
| Niemann-Pick C | *NPC* | G2864C | Persian | Somers 等 (2003) |
| Polycystic kidney disease[a] | *PKD1* | C10063A | Persian | Lyons 等 (2004) |
| Progressive retinal atrophy[a] | *CEP290* | IVS50 + 9T>G | Abyssinian | Menotti-Raymond 等 (2007) |
| Pyruvate kinase deficiency[a] | *PKLR* | 13 bp del in exon 6 | Abyssinian | (unpubl.)[b] |
| Spinal muscular atrophy[a] | *LIX1* | 140 kb del, exons 4–6 | Maine coon | Fyfe 等 (2006) |

a 表示在该品种中现在已经可以区分开的疾病；b 表示迄今没有被报道的突变。

个品种还能分离出两种神经节苷脂沉积症，GM1 和 GM2，同时该品种还具有点状毛色的表型，所有的这些都是品种中的不良特性。

猫类中两种最普遍的紊乱突变会引起心脏和肾脏的疾病。肥厚型心肌病（HCM）是一种异质的心脏疾病，该病在几个品种中均有发现，其中包括美国短毛猫、孟加拉猫、拉格道尔猫和斯芬克斯猫。然而，在科学记载中 HCM 只在缅因猫种存在可遗传性（Kittleson et al.　1999）。肌凝蛋白 C 结合蛋白（*MYCBP*）的突变与缅因猫和拉格道尔猫出现 HCM 疾病时的临床表现密切相关（Merus et al.　2005，2007），并且该突变被认为是引起该 HCM 疾病的成因突变。尽管在缅因猫品种中该疾病的发生频率仍没有被明确的调查出来，但是，一个为了减少缅因猫品种中该疾病发生的遗传分型计划已经在进行当中，这个计划结合运用了心回波图评估技术。在其他品种的 HCM 中也衡量评价了与这相同的突变，但是同人类的情况一样，HCM 在猫类中表现为异质性，而且在其他品种中该突变并不与疾病相关。

其他猫类疾病的发现，如波斯猫的多囊性肾病（PKD），都得益于一种结合了遗传连锁分析和候补基因方法的手段（Lyons et al.　2004，Young et al.　2005）。猫科动物的 PKD 是 1990 年首次作为一种常染色体显性遗传性特性被临床记载（Biller et al.　1990，1996）。全世界至多有 38％的波斯猫都患有 PKD（Beck and Lavelle　2001；Barrs et al.　2001；Cannon et al.　2001；Barthez et al.　2003），这就使得 PKD 成了家猫中最常见的遗传疾病。人类的 PKD 单在美国就影响了 600 000 人口，在世界范围则

有 12 500 000 位患者。因此患有 PKD 的患者数量要多于囊肿性纤维化、肌肉萎缩症、血友病、唐氏综合征和镰性细胞贫血症的患者总和。超过 90%的 PKD 是遗传性的。超过 60%的 PKD 患者会出现肾衰竭或者晚期肾脏疾病（ESRD）症状。猫类中，超过 95%患有 PKD 的猫在 8 个月大的时候就出现肾脏囊肿，该症状可以通过超音波来确诊。根据病情严重程度的不同，有些猫可以在患有 PKD 的情况下仍然活 10～14 年的正常寿命，有些则在发病后几年内就会死亡。人类中该疾病的突变和病情发展状况也是相类似的。利用超声波对肾囊肿进行检测的方法可以正确探查到患有 PKD 的波斯猫的家谱谱系。在人类中，*PKD1* 和 *PKD2* 两个基因与大部分 PKD 的发生有关系。这两个主要基因会产生非常大量的转录拷贝，同时，由于现在已经知道了部分的家猫基因序列，所以研究者先利用了连锁分析来寻找猫科动物 PKD 的候补基因而不是先通过候补基因方法（Young et al.　2005）。一旦某个猫类微卫星标记显示出与猫类 *PKD1* 基因组区域有明显的连锁，基因扫描就可以确定发生了一个精氨酸到 OPA 终止子的转换，这种转换大概会破坏 25%的多囊蛋白-1 的形成（Lyons et al.　2004）。

大多数猫类疾病和它们的突变都是针对于特殊的品种的。然而，在波斯猫和暹罗猫家族中发现的所有疾病都可以普及到其他品种中，因为这些品种会被当做改变形态结构产生新品种时所用到的基础品种来利用。例如，PKD 在苏格兰折耳猫、塞尔凯克卷毛猫（Selkirk Rex）和英国短毛家猫这些品种中都有记载表明存在，这些短头型品种在以前都利用过波斯猫来进行形态上的改良（Lyons et al.　2004）。HCM 被发现在具有非常庞大种群的缅因猫品种中非常普及，但是 GM1 和 GM2 这两种疾病只在一个小种群数量的品种中被发现并且以很低的频率出现。所以，猫品种培育者现在都强烈希望能够利用 DNA 检测的方法来对每种作为配种载体的猫进行检测，但是他们却正在和他们所培育品种的疾病管理判定方法进行对抗。

总的来说，23 个突变会导致猫类的疾病，4 个毛色位点会表现出额外的 7 个突变，同时一个血型的突变和猫类的 B 血型相关。由于 *Tas1r2* 甜味受体基因的假基因化，猫类甚至被认为缺少对甜食的喜爱（Li et al.　2005，2006）。除了那 277 个可以在猫类遗传的特性之外，现在已经开始了对于其他一些突变的鉴定和探究。活跃连锁研究已经发现了白色点状性状、虎斑状性状、橘色性状和视网膜萎缩性状的基因或者遗传区域，并且发现了缅甸猫存在的一个颅面缺陷。正在不断发展中的遗传手段和知识将会加快对于猫类单个基因性状的突变研究，并且让猫类研究成为对复杂性状和健康状况研究的助力。

## 猫科动物基因组学

早期对家猫的染色体带型研究揭示了一个非常明显的包括有 18 个常染色体对和 XY 性染色体对的染色体组型，补足了 38 条猫的染色体（Wurster-Hill and Gray　1973）。将染色体按字母顺序编组的传统编组方法是以染色体的大小和着丝粒的位置为基础的，就在最近研究者对这种传统编组方法进行了重命名，使之成为专属于猫类的更为标准化的术语（Cho et al.　1997）。原始的光学显微镜和吉姆萨带也表明家猫

的染色体结构具有高度的代表性，代表了所有猫科动物的染色体结构，甚至代表了食肉动物的染色体结构（Modi and O’Brien 1988）。在36个现存的猫科动物中只记录了有小染色体重排的现象。值得注意的是，南美洲豹家猫系的罗伯逊易位现象导致了染色体补足数的减少，出现了2N＝36的现象（Wurster-Hill and Gray 1973）。染色体大小的改变会有利于通过流式分类进行染色体涂染（Wienberg et al. 1997）。染色体涂染技术支持了早期的体细胞杂交图谱，因为猫类在染色体排列上和人类有着相当高的保守性，尤其是当和小鼠比较的时候（Stanyon et al. 1999）。所以，染色体涂染为猫类基因组组织方式提供了很好的描述手段，这极大地推动了候补基因方法，因为猫中特定基因的位点可以通过与人类遗传图谱进行比较的方法而预先考虑到。

猫的遗传图谱和放射性杂交图放大了那些由染色体研究所得来的低分辨率的遗传比对。孟加拉猫是由家猫（最初来自于阿比尼西亚猫和埃及猫或者 Indian Maus）和亚洲豹纹猫的不同亚种杂交而来的杂交后代。孟加拉猫这个品种是从1960年就开始出现的，尽管并不是所有的注册处都正式承认孟加拉猫，但是由于它独特的颜色和皮毛的样式，所以它现在已经是一种非常流行的猫类品种。孟加拉猫亲本类型的进化距离实际上是非常大的。所以，孟加拉猫的谱系是首个猫类基因重组图谱的基础（Menotti-Raymond et al. 1999）。孟加拉猫这个品种存在着很多健康问题和患病倾向，如HCM、视网膜变形和慢性肠炎性关节炎。它的这些情况对在杂种猫类种群中进行LD和混合型的绘图来说是非常好的条件。这种基于种间杂交基础上的连锁图谱通常能包括大概250个微卫星标记（Menotti-Raymond et al. 1999，2003b），这些微卫星标记对于初步研究表型特征分离的家族的连锁是非常有效的方法。在未来的几年里，连锁图谱应当会得到更新。现在，连锁图谱已经促进了定向的候补基因方法的发展，如对 *PKD* 的研究（Young et al. 2005）、对虎斑猫的连锁分析（Lyons et al. 2006）、对白色斑点性状的研究（Cooper et al. 2006）和对橘色性状的研究（Grahn et al. 2005）。遗传图谱同时也引起了首个利用定位克隆而进行的疾病基因分离，如 *LIX1* 基因，*LIX1* 的异常会导致缅因猫脊髓性肌萎缩的发生（He et al. 2005；Fyfe et al. 2006）。迄今所研究出来的5000个猫的Rad放射性杂交图中总共包括了1784个标记物（Murphy et al. 1999，2000，2006；Menotti-Raymond et al. 2003a），这些图谱通过和人类进行比较支持了保守基因组组成的作用，并且协助了序列重叠群的建立。猫对人类健康的重要性、比较基因组学和进化研究这三个方面的意义都促使美国国立卫生研究院-美国国家人类基因组研究计划协会（NIH-NHGRI）决定建成一个低范围、2X的猫类基因组序列。在布罗德研究中心和 AgenCourt 的带领下，已经从高度近亲交配的阿比西尼亚猫的序列中鉴定出了327037个SNP。同一个品种的另一个7X范围的序列也已经被提上议程并马上完成，这将在未来为研究者提供一个范围更深入的序列草图。

对猫类进行的LD评估尚在进行当中还没有公布。由于家猫品种的动态性，所以很可能猫的LD会比狗的范围要小，但是也会比人类的更为广泛和多方面。现在已经发展出了一个经过国际检测的基于微卫星基础上的DNA概要分析，这个概要分析是为了进行家猫血统和个体鉴定而产生出来，结果表明在分析家猫时需要比其他物种更少的标记物，这是因为大部分猫类品种在所有的标记物上都具有足够的变化。在这个概要分析的

发展中囊括了19个微卫星标记，而且这19个标记都已经在很多家猫DNA检测样品中进行了基因型的确定。大多数标记都是由二核苷酸组成的。除了常染色体标记物以外，该分析包括了两个性别特异性的标记物，*amelogenin*和锌指XY，它们都是为X和Y染色体产生基因型的。和其他物种所用到的分析相比，国际猫类DNA概要分析具有排他性的特点，随品种的不同其范围在90.08%～99.79%浮动，在随机交配品系猫类种群中浮动范围在99.47%～99.87%。然而，要获得足够单一的概率只需要有10个标记即可（表30-4）。相比之下，狗类品种和其他物种都需要15个甚至更多的标记物才能获得足够单一的概率。

## 结论

对于猫类的遗传研究和资讯在过去的几十年间得到了飞速的发展，并且现在仍然有着继续飞跃式发展的可能。现在，猫类的神秘之处正在一点点地被揭开，但是猫类的原始祖先和引起多种疾病和特性的突变仍没有被鉴定出来。迄今为止的谱图、更深一层的基因组序列、细菌人工合成染色体和其他方面的文库都可以在各个领域为将来的遗传研究提供高效和充足的资讯。猫是研究传染病和获得性疾病的良好模型，同时它们在形态和行为上的特点也在不同的品种和种群中有所不同。不断进步的兽医学设备和药品都支持利用家猫建立新模型的观点，而且，现在已经存在的模型也对基因治疗和干细胞研究非常有用，这些现存的模型打开了人类和猫科动物健康与药物研究的进步之门。

## 致谢

该研究计划的基金由NIH-NCRP提供给L. A. Lyons，基金的资助编号为RR016094。资助本研究计划的还有：Winn Feline Foundation、George and Phyllis Miller Feline Health Fund、Center for Companion Animal Health、Koret Center for Veterinary Genetics、School of Veterinary Medicine和UC. Davis。

**表 30-4 为进行猫类亲缘分析和鉴定而被选作为“核心”分析的遗传标记**

| 标记物 | 杂色体 | 重复 | 正向引物 5′-3′<br>反向引物 5′-3′ | 标记 | μM | PE(Min-Max)(品种) | PE(Min-Max)(随机) |
|---|---|---|---|---|---|---|---|
| FCA069 | B4 | AC | AATCACTCATGCACGAATGC<br>AATTTAACGTTAGGCTTTTTGCC | VIC | 0.20 | 0.1324–0.5336 | 0.3958–0.5948 |
| FCA075 | E2 | TG | ATGCTAATCAGTGGCATTTGG<br>GAACAAAAATTCCAGACGTGC | NED | 0.10 | 0.1442–0.5771 | 0.4240–0.5992 |
| FCA105 | A2 | TG | TTGACCCTCATACCTTCTTTGG<br>TGGGAGAATAAATTTGCAAAGC | PET | 0.20 | 0.2221–0.5585 | 0.6110–0.7101 |
| FCA149[a] | B1 | TG | CCTATCAAAGTTCTCACCAAATCA<br>GTCTCACCATGTGTGGGATG | PET | 0.18 | 0.1783–0.5995 | 0.3586–0.5767 |
| FCA220 | F2 | CA | CGATGGAAATTGTATCCATGG<br>GAATGAAGGCAGTCACAAACTG | FAM | 0.30 | 0.0000–0.3383 | 0.1851–0.4221 |
| FCA229 | A1 | GT | CAAACTGACAAGCTTAGAGGGC<br>GCAGAAGTCCAATCTCAAAGTC | NED | 0.25 | 0.0452–0.5131 | 0.3927–0.5813 |
| FCA310[a] | C2 | $(CA)_5TA(CA)_7TA(CA)_8$ | TTAATTGTATCCCAAGTGGTCA<br>TAATGCTGCAATGTAGGGCA | FAM | 0.30 | 0.1196–0.5256 | 0.3417–0.5611 |
| FCA441[b] | D3 | TAGA | ATCGGTAGGTAGGTAGATATAG<br>GCTTGCTTCAAAATTTTCAC | VIC | 0.15 | 0.2061–0.5774 | 0.3388–0.5505 |
| FCA678[c] | A1 | AC | TCCCTCAGCAATCTCCAGAA<br>GAGGGAGCTAGCTGAAATTGTT | NED | 0.25 | 0.0415–0.4908 | 0.3016–0.5715 |
| AMEL[d] | XY | — | CGAGGTAATTTTTCTGTTTACT<br>GAAACTGAGTCAGAGAGGC | | | n.a. | n.a. |
| ZFXY[d] | XY | — | AAGTTTACACAACCACCTGG<br>CACAGAATTTACACTTGTGCA | PET | 0.20 | n.a. | n.a. |
| | | | | Total PE | | 0.9008–0.9979 | 0.9947–0.9987 |

注：n. a 表示不可用。

a 表示前 10 个发表的猫科动物微卫星的标记物；b 表示现在尚处于可辩论状态的标记物；c 表示两个在 X 和 Y 染色体上的标记物，它们在经过对比试验后被加入分析之中；d 表示为 FCA678 新设计出来的引物，它们可以产生一个比已发表的引物少 30 个 bp 的产物。

# 参考文献


Barrs V.R., Gunew M., Foster S.F., Beatty J.A., and Malik R. 2001. Prevalence of autosomal dominant polycystic kidney disease in Persian cats and related-breeds in Sydney and Brisbane. *Aust. Vet. J.* **79:** 257–259.

Barthez P.Y., Rivier P., and Begon D. 2003. Prevalence of polycystic kidney disease in Persian and Persian related cats in France. *J. Feline Med. Surg.* **5:** 345–347.

Beck C. and Lavelle R.B. 2001. Feline polycystic kidney disease in Persian and other cats: A prospective study using ultrasonography. *Aust. Vet. J.* **79:** 181–184.

Berg T., Tollersrud O.K., Walkley S.U., Siegel D., and Nilssen O. 1997. Purification of feline lysosomal α-mannosidase, determination of its cDNA sequence and identification of a mutation causing α-mannosidosis in Persian cats. *Biochem. J.* **328:** 863–870.

Bighignoli B., Grahn R.A., Millon L.V., Longeri M., Polli M., and Lyons L.A. 2007. Genetic mutations for the feline AB blood group identified in *CMAH*. *BMC Genet.* **8:** 27.

Biller D.S., Chew D.J., and DiBartola S.P. 1990. Polycystic kidney disease in a family of Persian cats. *J. Am. Vet. Med. Assoc.* **196:** 1288–1290.

Biller D.S., DiBartola S.P., Eaton K.A., Pflueger S., Wellman M.L., and Radin M.J. 1996. Inheritance of polycystic kidney disease in Persian cats. *J. Hered.* **87:** 1–5.

Cannon M.J., MacKay A.D., Barr F.J., Rudorf H., Bradley K.J., and Gruffydd-Jones T.J. 2001. Prevalence of polycystic kidney disease in Persian cats in the United Kingdom. *Vet. Rec.* **149:** 409–411.

Cavalli-Sforza L.L. and Edwards A.W.F. 1967. Phylogenetic analysis: Models and estimation procedures. *Evolution* **21:** 550–570.

Cho K.W., Youn H.Y., Watari T., Tsujimoto H., Hasegawa A., and Satoh H. 1997. A proposed nomenclature of the domestic cat karyotype. *Cytogenet. Cell Genet.* **79:** 71–78.

Cooper M.P., Fretwell N., Bailey S.J., and Lyons L.A. 2006. White spotting in the domestic cat (*Felis catus*) maps near *KIT* on feline chromosome B1. *Anim. Genet.* **37:** 163–165.

Crawley A.C., Yogalingam G., Muller V.J., and Hopwood J.J. 1998. Two mutations within a feline mucopolysaccharidosis type VI colony cause three different clinical phenotypes. *J. Clin. Invest.* **101:** 109–119.

Drogemuller C., Rufenacht S., Wichert B., and Leeb T. 2007. Mutations within the FGF5 gene are associated with hair length in cats. *Anim. Genet.* **38:** 218–221.

De Maria R., Divari S., Bo S., Sonnio S., Lotti D., Capucchio M.T., and Castagnaro M. 1998. β-galactosidase deficiency in a Korat cat: A new form of feline $G_{M1}$-gangliosidosis. *Acta Neuropathol.* **96:** 307–314.

Eizirik E., Yuhki N., Johnson W.E., Menotti-Raymond M., Hannah S.S., and O'Brien S.J. 2003. Molecular genetics and evolution of melanism in the cat family. *Curr. Biol.* **13:** 448–453.

Felsenstein J. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164–166.

Fogle B. 1997. *The encyclopedia of the cat.* DK Publishing, New York.

Fyfe J.C., Kurzhals R.L., Lassaline M.E., Henthorn P.S., Alur P.R., Wang P., Wolfe J.H., Giger U., Haskins M.E., Patterson D.F, et al. 1999. Molecular basis of feline β-glucuronidase deficiency: An animal model of mucopolysaccharidosis VII. *Genomics* **58:** 121–128.

Fyfe J.C., Menotti-Raymond M., David V.A., Brichta L., Schaffer A.A., Agarwala R., Murphy W.J., Wedemeyer W.J., Gregory B.L., Buzzell B.G., et al. 2006. An ~140-kb deletion associated with feline spinal muscular atrophy implies an essential *LIX1* function for motor neuron survival. *Genome Res.* **16:** 1084–1090.

Fyfe J.C., Kurzhals R.L., Hawkins M.G., Wang P., Yuhki N., Giger U., Van Winkle T.J., Haskins M.E., Patterson D.F., and Henthorn P.S. 2007. A complex rearrangement in GBE1 causes both perinatal hypoglycemic collapse and late-juvenile-onset neuromuscular degeneration in glycogen storage disease type IV of Norwegian forest cats. *Mol. Genet. Metab.* **90:** 383–392.

Giger U., Tcherneva E., Caverly J., Seng A., Huff A. M., Cullen K., Van Hoeven M., Mazrier H., and Haskin M.E. 2006. A missense point mutation in *N*-acetylglucosamine-1-phosphotransferase causes mucolipidosis II in domestic shorthair cats. *J. Vet. Intern. Med.* **20:** 781.

Ginzinger D.G., Lewis M.E., Ma Y., Jones B.R., Liu G., and Jones S.D. 1996. A mutation in the lipoprotein lipase gene is the molecular basis of chylomicronemia in a colony of domestic cats. *J. Clin. Invest.* **97:** 1257–1266.

Goree M., Catalfamo J.L., Aber S., and Boudreaux M.K. 2005. Characterization of the mutations causing hemophilia B in 2 domestic cats. *J. Vet. Intern. Med.* **19:** 200–204.

Grahn R.A., Lemesch B.M., Millon L.V., Matise T., Rogers Q.R., Morris J.G., Fretwell N., Bailey S.J., Batt R.M., and Lyons L.A. 2005. Localizing the X-linked orange colour phenotype using feline resource families. *Anim. Genet.* **36:** 67–70.

He Q., Lowrie C., Shelton G.D., Castellani R.J., Menotti-Raymond M., Murphy W., O'Brien S.J., Swanson W.F., and Fyfe J.C. 2005. Inherited motor neuron disease in domestic cats: A model of spinal muscular atrophy. *Pediatr. Res.* **57:** 324–330.

He X., Li C.M., Simonaro C.M., Wan Q., Haskins M.E., Desnick R.J., and Schuchman E.H. 1999. Identification and characterization of the molecular lesion causing mucopolysaccharidosis type I in cats. *Mol. Genet. Metab.* **67:** 106–112.

Ibsen H.L. 1916. Tricolor inheritance. III. Tortoiseshell cats. *Genetics* **1:** 377–386.

Imes D.L., Geary L.A., Grahn R.A., and Lyons L.A. 2006. Albinism in the domestic cat (*Felis catus*) is associated with a *tyrosinase* (*TYR*) mutation. *Anim. Genet.* **37:** 175–178.

Ishida Y., David V.A., Eizirik E., Schaffer A.A., Neelam B.A., Roelke M.E., Hannah S.S., O'Brien S.J., and Menotti-Raymond M. 2006. A homozygous single-base deletion in *MLPH* causes the *dilute* coat color phenotype in the domestic cat. *Genomics* **88:** 698–705.

Johnson W.E., Eizirik E., Pecon-Slattery J., Murphy W.J., Antunes A., Teeling E., and O'Brien S.J. 2006. The late Miocene radiation of modern Felidae: A genetic assessment. *Science* **311:** 73–77.

Kanae Y., Endoh D., Yamato O., Hayashi D., Matsunaga S., Ogawa H., Maede Y., and Hayashi M. 2006. Nonsense mutation of feline β-hexosaminidase β-subunit (*HEXB*) gene causing Sandhoff disease in a family of Japanese domestic cats. *Res. Vet. Sci.* **82:** 54–60.

Kitchener A. 1991. *The natural history of wild cats.* Cornell University Press, New York.

Kittleson M.D., Meurs K.M., Munro M.J., Kittleson J.A., Liu S.K., Pion P.D., and Towbin J.A. 1999. Familial hypertrophic cardiomyopathy in Maine coon cats: An animal model of human disease. *Circulation* **99:** 3172–3180.

Li X., Li W., Wang H., Bayley D.L., Cao J., Reed D.R., Bachmanov A.A., Huang L., Legrand-Defretin V., Beauchamp G.K., and Brand J.G. 2006. Cats lack a sweet taste receptor. *J. Nutr.* **136:** 1932S–1934S.

Li X., Li W., Wang H., Cao J., Maehashi K., Huang L., Bachmanov A.A., Reed D.R., Legrand-Defretin V., Beauchamp G.K., and Brand J.G. 2005. Pseudogenization of a sweet-receptor gene accounts for cats' indifference toward sugar. *PLoS Genet.* **1:** 27–35.

Lipinski M.J., Amigues Y., Blasi M., Broad T.E., Cherbonnel C., Cho G.J., Corley S., Daftari P., Delattre D.R., Dileanis S., et al. 2007. An international microsatellite-based DNA profiling panel for the domestic cat (*Felis catus*). *Anim. Genet.* (in press).

Louwerens M., London C.A., Pedersen N.C., and Lyons L.A. 2005. Feline lymphoma in the post-feline leukemia virus era. *J. Vet.*

*Intern. Med.* **19:** 329–335.

Lyons L.A., Foe I.T., Rah H.C., and Grahn R.A. 2005a. Chocolate coated cats: *TYRP1* mutations for brown color in domestic cats. *Mamm. Genome* **16:** 356–366.

Lyons L.A., Imes D.L., Rah H.C., and Grahn R.A. 2005b. Tyrosinase mutations associated with Siamese and Burmese patterns in the domestic cat (*Felis catus*). *Anim. Genet.* **36:** 119–126.

Lyons L.A., Biller D.S., Erdman C.A., Lipinski M.J., Young A.E., Roe B.A., Qin B., and Grahn R.A. 2004. Feline polycystic kidney disease mutation identified in PKD1. *J. Am. Soc. Nephrol.* **15:** 2548–2555.

Lyons L.A., Bailey S.J., Baysac K.C., Byrns G., Erdman C.A., Fretwell N., Froenicke L., Gazlay K.W., Geary L.A., Grahn J.C., et al. 2006. The Tabby cat locus maps to feline chromosome B1. *Anim. Genet.* **37:** 383–386.

Malek J. 2006. *The cat in ancient Egypt*. British Museum Press, London.

Martin D.R., Krum B.K., Varadarajan G.S., Hathcock T.L., Smith B.F., and Baker H.J. 2004. An inversion of 25 base pairs causes feline $G_{M2}$ gangliosidosis variant 0. *Exp. Neurol.* **187:** 30–37.

Martin D.R., Cox N.R., Morrison N.E., Kennamer D.M., Peck S.L., Dodson A.N., Gentry A.S., Griffin B., Rolsma M.D., and Baker H.J. 2005. Mutation of the $G_{M2}$ activator protein in a feline model of $G_{M2}$ gangliosidosis. *Acta Neuropathol.* **110:** 443–550.

Menotti-Raymond M.A. and O'Brien S.J. 1995. Evolutionary conservation of ten microsatellite loci in four species of Felidae. *J. Hered.* **86:** 319–322.

Menotti-Raymond M.A., David V.A., Wachter L.L., Butler J.M., and O'Brien S.J. 2005. An STR forensic typing system for genetic individualization of domestic cat (*Felis catus*) samples. *J. Forensic Sci.* **50:** 1061–1070.

Menotti-Raymond M., David V.A., Agarwala R., Schaffer A.A., Stephens R., O'Brien S.J., and Murphy W.J. 2003a. Radiation hybrid mapping of 304 novel microsatellites in the domestic cat genome. *Cytogenet. Genome Res.* **102:** 272–276.

Menotti-Raymond M., David V.A., Lyons L.A., Schaffer A.A., Tomlin J.F., Hutton M.K., and O'Brien S.J. 1999. A genetic linkage map of microsatellites in the domestic cat (*Felis catus*). *Genomics* **57:** 9–23.

Menotti-Raymond M., David V.A., Schaffer A.A., Stephens R., Wells D., Kumar-Singh R., O'Brien S.J., and Narfstrom K. 2007. Mutation in CEP290 discovered for cat model of human retinal degeneration. *J. Hered.* (in press).

Menotti-Raymond M., David V.A., Chen Z.Q., Menotti K.A., Sun S., Schaffer A.A., Agarwala R., Tomlin J.F., O'Brien S.J., and Murphy W.J. 2003b. Second-generation integrated genetic linkage/radiation hybrid maps of the domestic cat (*Felis catus*). *J. Hered.* **94:** 95–106.

Meurs K.M., Norgard M.M., Ederer M.M., Hendrix K.P., and Kittleson M.D. 2007. A substitution mutation in the myosin binding protein C gene in ragdoll hypertrophic cardiomyopathy. *Genomics* (in press).

Meurs K.M., Sanchez X., David R.M., Bowles N.E., Towbin J.A., Reiser P.J., Kittleson J.A., Munro M.J., Dryburgh K., Macdonald K.A., and Kittleson M.D. 2005. A cardiac myosin binding protein C mutation in the Maine Coon cat with familial hypertrophic cardiomyopathy. *Hum. Mol. Genet.* **14:** 3587–3593.

Modi W.S. and O'Brien S.J. 1988. Quantitative cladistic analysis of chromosomal banding data among species in three orders of mammals: Hominoid primates, felids, and arvicolid rodents. *In Chromosome structure and function* (Eds. J. Perry Gustafson and R. Appels). Plenum Publishing, New York.

Morris D. 1999. Cat breeds of the world: A complete illustrated encyclopedia. Viking Penguin, New York.

Muldoon L.L., Neuwelt E.A., Pagel M.A., and Weiss D.L. 1994. Characterization of the molecular defect in a feline model for type II $G_{M2}$-gangliosidosis (Sandhoff disease). *Am. J. Pathol.* **144:** 1109–1118.

Murphy W.J., Menotti-Raymond M., Lyons L.A., Thompson M.A., and O'Brien S.J. 1999. Development of a feline whole genome radiation hybrid panel and comparative mapping of human chromosome 12 and 22 loci. *Genomics* **57:** 1–8.

Murphy W.J., Sun S., Chen Z., Yuhki N., Hirschmann D., Menotti-Raymond M., and O'Brien S.J. 2000. A radiation hybrid map of the cat genome: Implications for comparative mapping. *Genome Res.* **10:** 691–702.

Murphy W.J., Davis B., David V.A., Agarwala R., Schaffer A.A., Pearks Wilkerson A.J., Neelam B., O'Brien S J., and Menotti-Raymond M. 2006. A 1.5-Mb-resolution radiation hybrid map of the cat genome and comparative analysis with the canine and human genomes. *Genomics* **89:** 189–196.

Nei M. 1972. Genetic distance between populations. *Am. Naturalist* **106:** 283–292.

O'Brien S.J., Haskins M.E., Winkler C.A., Nash W.G., and Patterson D.F. 1986. Chromosomal mapping of beta-globin and albino loci in the domestic cat. A conserved mammalian chromosome group. *J. Hered.* **77:** 374–378.

Pilgrim K.L., McKelvey K.S., Riddle A.E., and Schwartz M.K. 2005. Felid sex identification based on noninvasive genetic samples. *Mol. Biol. Notes* **5:** 60–61.

Pritchard J.K., Stephens M., and Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

Seidensticker J. and Lumpkin S. 1991. *Great cats: Majestic creatures of the wild*. Rodale Press, Emmaus, Pennsylvania.

Somers K.L., Royals M.A., Carstea E.D., Rafi M.A., Wenger D.A., and Thrall M.A. 2003. Mutation analysis of feline Niemann-Pick C1 disease. *Mol. Genet. Metab.* **79:** 99–103.

Stanyon R., Yang F., Cavagna P., O'Brien P.C., Bagga M., Ferguson-Smith M.A., and Wienberg J. 1999. Reciprocal chromosome painting shows that genomic rearrangement between rat and mouse proceeds ten times faster than between humans and cats. *Cytogenet. Cell Genet.* **84:** 150–155.

Sunquist M. and Sunquist F. 2002. *Wild cats of the world*. University of Chicago Press, Illinois.

Wienberg J., Stanyon R., Nash W.G., O'Brien P.C., Yang F., O'Brien S.J., and Ferguson-Smith M.A. 1997. Conservation of human vs. feline genome organization revealed by reciprocal chromosome painting. *Cytogenet. Cell Genet.* **77:** 211–217.

Winand N.J., Edwards M., Pradhan D., Berian C.A., and Cooper B.J. 1994. Deletion of the dystrophin muscle promoter in feline muscular dystrophy. *Neuromuscul. Disord.* **4:** 433–445.

Wurster-Hill D.H. and Gray C.W. 1973. Giemsa banding patterns in the chromosomes of twelve species of cats (Felidae). *Cytogenet. Cell Genet.* **12:** 388–397.

Yogalingam G., Hopwood J.J., Crawley A., and Anson D.S. 1998. Mild feline mucopolysaccharidosis type VI. Identification of an *N*-acetylgalactosamine-4-sulfatase mutation causing instability and increased specific activity. *J. Biol. Chem.* **273:** 13421–13429.

Yogalingam G., Litjens T., Bielicki J., Crawley A.C., Muller V., Anson D.S., and Hopwood J.J. 1996. Feline mucopolysaccharidosis type VI. Characterization of recombinant *N*-acetylgalactosamine 4-sulfatase and identification of a mutation causing the disease. *J. Biol. Chem.* **271:** 27259–27265.

Young A.E., Biller D.S., Herrgesell E.J., Roberts H.R., and Lyons L.A. 2005. Feline polycystic kidney disease is linked to the *PKD1* region. *Mamm. Genome* **16:** 59–65.

Zohary D. and Hopf M. 2002. Domestication of plants in the old world. 3 ed. Oxford University Press, Oxford.

# 31 狗

Kerstin Lindblad-Toh[1] and Elaine A. Ostrander[2]

*[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142; [2]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892*



## 简介

在过去的 5 年中，犬类（canine）遗传学和基因组学领域有了巨大的发展。驯养狗的独特繁殖历史，包括在过去几百年中产生的数以百计的不同品种，再结合一套新的基因组工具，使犬科基因组研究走在了基因组学领域的前列。现有资源包括一个完整的基因组序列、一个 2 500 000 的 SNP 图谱、对单体型结构与品种之间关系不断深入的了解及 SSLP 和 SNP 作图工具。本章，我们讨论了这些新进展和遗传变异在控制感兴趣的表型（如孟德尔疾病）和复杂性状中所起的作用。我们预测在接下来的几年中会作出多种疾病和其他复杂性状的图谱，其中一些会有助于解决人类健康问题和生物学问题。

### 狗的历史和品种

驯养狗是从家庭犬科动物进化而来的最新种（Wayne et al. 1987a，b，1997），和狼样犬科动物，如灰狼、郊狼和豺狗，有着共同的进化枝（图 31-1）。人们认为狗起源

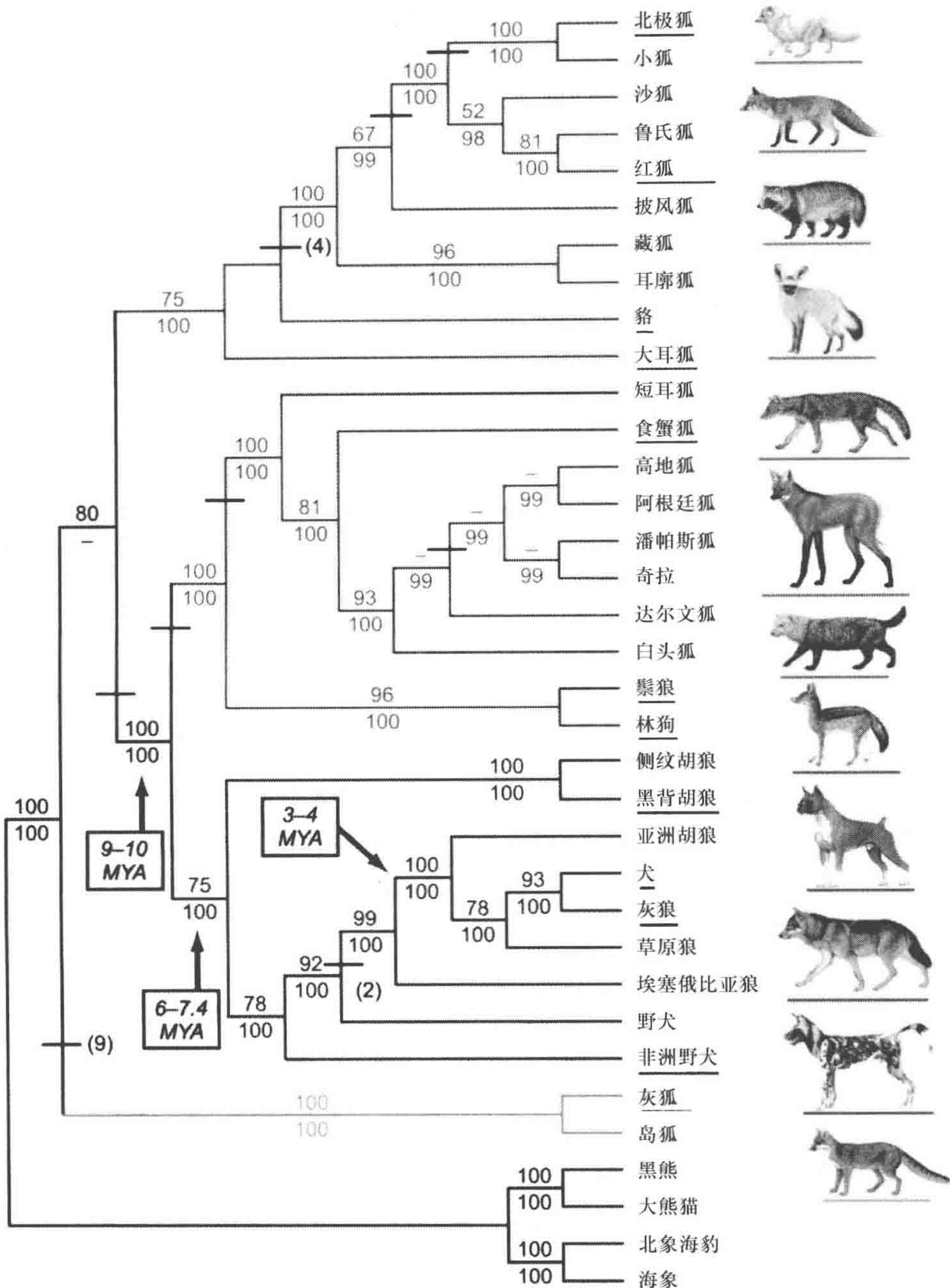



图 31-1 犬科树。犬科系统发生树是根据约 15 kb 的外显子和内含子序列作出的。此树通过使用最大简易性作为最优标准而得以构建出来，是单个最简树。分支的颜色标记了红狐狸样进化枝（红色）、南美进化枝（绿色）、狼样进化枝（蓝色）和灰/岛狐狸进化枝（橘黄色）。下划线处为示例种。树根由黑熊、大熊猫、象海豹和海象作出（授权复印自 Lindblad-Toh 等 2005）。（见图版）

于 30 000 年以前，最初的驯化事件发生在东亚（Vila et al. 1997；Sauolaine et al. 2002；Lindblad-Toh et al. 2005）。人们相信大部分的狗品种是紧随着这些主要的驯化事件产生于过去的 200～300 年中。结果，现在有超过 400 种驯养的狗品种，它们有着很大的生理和行为差异。在美国，AKC 登记了 155 个品种。

术语“品种”是狗类爱好者和遗传学家提出的。要确定狗属于一个品种，AKC 要求狗的双亲是同一个品种中已记录的成员。这样，现代狗就变成了封闭的繁殖种群，很少引入新等位基因。结果，狗类品种就具有较低水平的遗传异质性（1SNP/1600bp），而在混合品种中的遗传异质性则较高（1SNP/900bp）（Lindblad-Toh et al. 2005）。这种现象由许多影响因素造成，这些因素包括少量决定了一些品种的奠基者、种群瓶颈和交配库中一些雄性领导者的特权过多。因此美国现有的大约 1000 万只纯种狗是研究种群遗传学和遗传变异在控制简单和复杂性状中的作用的理想群体。

一些研究解决了品种定义中存在的问题（Koskinen and Bredbacka 2000；Koskinen 2003；Parker et al. 2004）。Parker 等（2004）对狗品种进行了深入的研究，他们利用来自 414 条狗上的遍及所有常染色体的 96（CA）n 重复单位的 SSLP 数据，使用聚类算法确定了一只狗属于哪个品种（图 31-2）（Park et al. 2004）。我们根据等位基因的相似模式将 85 个品种分成了 4 个聚类，大致代表了共同的祖先库。正在进行的研究包括更多的品种，这些结果对人类种群遗传多样性研究具有很大的意义。整体水平的核苷酸多态性水平是 $8\times10^{-4}$，与报道的人类中的水平差别不大。尽管如此，狗类品种间的遗传变异比人类种群中观察到的大（27.5% vs 5.4% by ANOVA）（Parker et al. 2004）。因此，狗品种是统计学意义上的实体。

驯养狗的这种种群结构使得它被用于单基因的和复杂的性状的研究，特别是那些与疾病易感性相关的性状的研究之中。在人类和伴侣动物中都存在的疾病和在人类种群中难以处理的疾病特别适合进行犬科作图研究。比较好的候选疾病包括癌症、糖尿病、自身免疫性疾病、运动神经疾病、聋、癫痫和心脏病（Ostrander and Friedrichsen 2004；Ostrander et al. 2004）。作为强烈的种群瓶颈导致品种产生和品种内杂交的一个结果，一些品种患某种特定疾病的比率较高，表明它的遗传风险因子含量丰富（Patterson et al. 1988；and Kruglyak 2000；Sargan 2004；Parker and Ostrander 2005）。

一个品种内低水平的变异也可能导致特定风险因子的固定，从而降低了噪声的整体水平，更容易观察到其他风险因子的效应。此外，狗的家庭历史记录完整，并且每年在狗的医学检查和健康护理资金的投入上仅次于人类（Patterson 2000；American Veterinary Medical Association 2002）。这样，在狗中已有数百种遗传病症得到了描述，仅次于在人类种群中的研究（Patterson et al. 1982；Patterson 2000；Sargan 2004）。这些都被整理在了一个名为 IDID 的在线数据库中，它的组织方式与在线人类孟德尔遗传（OMIM）相似（Sargan 2004）。而且，狗的生存环境大部分都与人类相同。总之，这些原因使得狗成为了人类遗传学和伴侣动物健康研究中感兴趣的单个性状和复杂性状进行遗传学研究的理想系统。表 31-1 总结了驯养狗的主要疾病和有患病风险的品种。

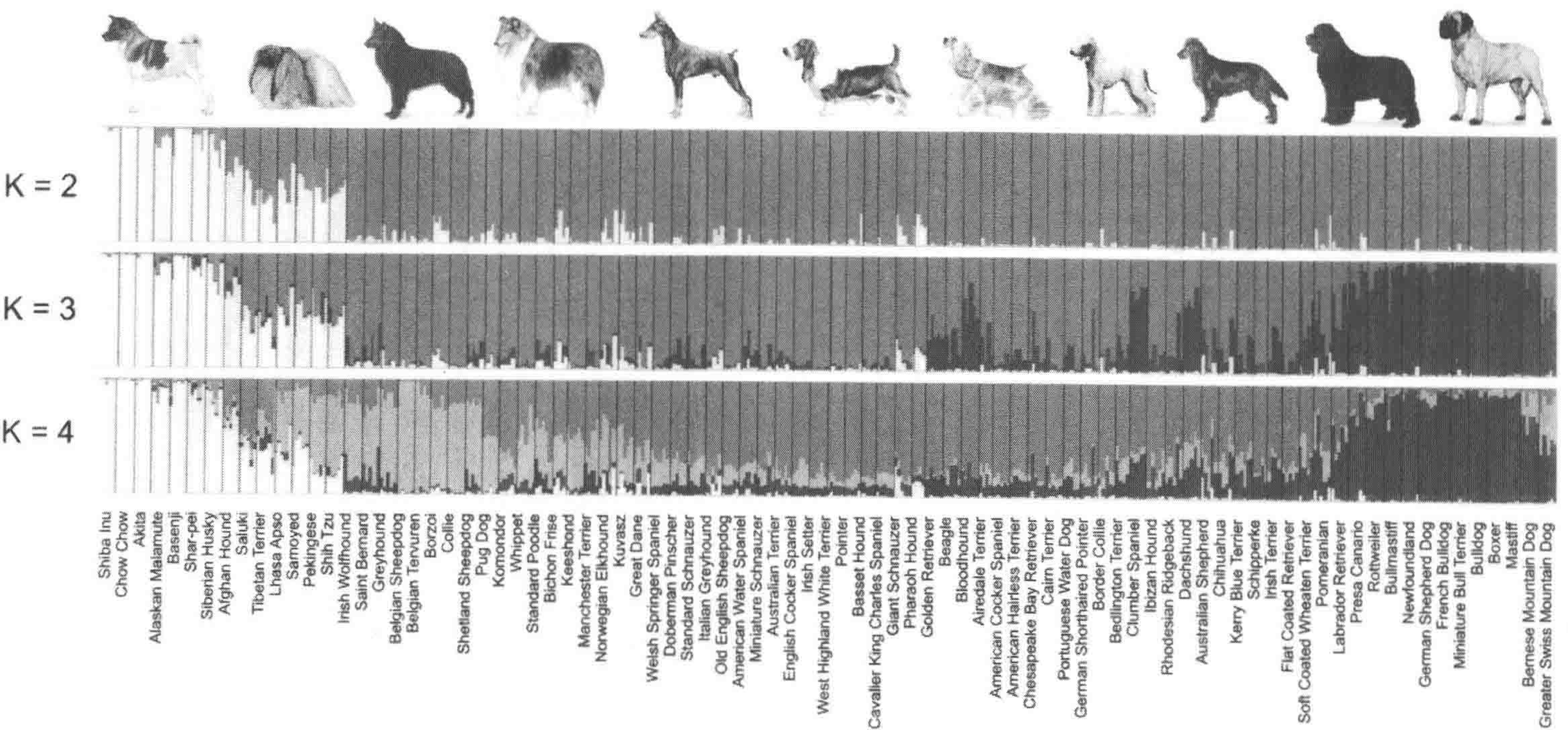



图31-2 驯养狗的种群结构。本图来自Parker及其同事的工作(2004)，他们使用85个碱基重复单位的微卫星序列得到了85个品种中每种5只狗的基因型。标记密度为30Mb，遍及所有的常染色体。使用计算机程序structure对其进行了分析，如果最初仅存在两个组(K=2)，此程序就可以将样本分成最优群体。在此水平上，它们选择了一组主要存在于亚洲或者源自古代的品种，因为它们与其余部分最不相关。这包括拉萨犬、沙皮犬和秋田犬。增加一个自由度(K=3)分出了獒组，包括拳师狗、斗牛犬和加纳利犬。在K=4时，鉴定出了一群工作犬品种，如圣伯纳犬和比利时牧羊犬。其余的狗，称为组4，主要为视力型嗅猎犬，包括斯班尼犬和嗅猎犬。每组狗彼此之间比其他的更加密切相关。每个品种的狗数量越多、品种数越多，分析就越精确

**表 31-1　驯养狗和品种的主要疾病类型**

| 疾病 | 高疾病风险品种 |
| --- | --- |
| 癌症 | 艾尔谷绠、日本秋田狗等 |
| 癫痫 | 澳洲绠、比利时玛利诺犬等 |
| 髋关节发育异常 | 艾尔谷绠、美国爱斯基摩犬等 |
| 甲状腺疾病 | 日本秋田狗、美国爱斯基摩犬等 |
| 过敏症 | 艾尔谷绠、美国水猎犬等 |
| 胃气胀 | 日本秋田狗、血猎犬等 |
| 心脏病 | 美国水猎犬、比利时玛伦牧羊犬等 |
| 自身免疫性疾病 | 艾尔谷绠、日本秋田犬等 |
| 进行性视网膜萎缩 | 艾尔谷梗、美国爱斯基摩犬等 |
| 白内障 | 贝林登绠、卷毛比雄犬等 |

在可以利用犬类基因组序列之前，遗传学家使用大的犬科家系来克服人类遗传学家所面临的一些不利因素。早期在狗中成功进行的遗传作图研究包括进行性视网膜萎缩（PRA）（Acland et al. 1994，1998，1999；Sidjanin et al. 2002）、铜中毒（Yuzbasiyan-Gurkan et al. 1997）、肾癌（Jonasdottir et al. 2000）和嗜睡症（Mignot et al. 1991；Lin et al. 1999）。所有这些研究都涉及高渗透性的、孟德尔性状的、大的、多基因家族，能用 SSLP 进行连锁分析。

## 犬类基因型序列

虽然犬科遗传学在过去几年中有了全面、显著的进展，但真正能大大提高犬科遗传学家解决感兴趣问题的能力的，却是一个覆盖了一只雌性拳师狗的大约 99%的常染色质基因组的一个 7.5×装配序列（assembly）。狗的常染色质基因组大约为 24 亿个碱基对，并且与人类基因组相比包含了 243 个保守的线性片段。在 19 000 个犬科基因中有 75%在人、老鼠和狗的同源物中具有 1-1-1 的对应关系。在一些网站上可以访问全基因组序列，这些网站有 http://www.genome.ucsc.edu;http://www.ncbi.nih.gov 和 http://www.ensembl.org。要得到更完整的犬科基因和图谱资源列表，见表 31-2。扩增犬类基因组序列的引物现在也已公开（O' Rourke 2005）。

虽然这个基因组序列精度很高并且连续性也很好（有一半的序列位于至少 180kb 长的连续的序列块中），但一小部分基因组还存在缺口或者结构不完整的区域（标记为“无保障区域”，在网站 http://www.broad.nit.edu/ftp/pub/assemblies/mammals/dog/canFam2可利用这些序列）。MIT 的 Broad 研究所和哈佛大学正在完善这些序列，它的目标是使部分的基因组达到完美的标准，以确定大多数较大的缺口和无保障区域，并且改正基因序列中的错误。要完成的工作包括 ENCODE 区域（http://www.genome.gov/12513456），这是代表随机的 1%基因组的 44 个 0.5～2.0 Mb 的区域，它们被用于注释人类基因组。这些基因组完善工作将会在 2007 年末得到一个几乎完整的基因组图谱。

**表 31-2 犬科基因组和作图资源**

| | |
|---|---|
| 基因组组装及浏览： | |
| NCBI | www.ncbi.nih.gov/Genbank |
| UCSC browser | www.genome.ucsc.edu |
| Ensembl | www.ensembl.org/index.html |
| Broad Institute | www.broad.mit.edu/ftp/pub/assemblies/mammals/dog/canFam2/ |
| Dog paper Supp Info | www.broad.mit.edu/ftp/pub/papers/dog_genome/suppinfo/ |
| 作图： | |
| RH map | www-recomgen.univ-rennes1.fr/Dogs/paper/FISH-RHmap.html |
| FISH map | www.cvm.ncsu.edu/mbs/breen_matthew.htm |
| SNP map | www.broad.mit.edu/ftp/pub/papers/dog_genome/snps_canfam2/ |
| 繁殖，关系及疾病： | |
| American Kennel Club | www.akc.org |
| Parker breed relationships | www.fhcrc.org/science/dog_genome/dog.html |
| IDID | www.vet.cam.ac.uk/idid/ |
| 作图工具： | |
| SSLP | www.cvm.tamu.edu/cgr/multiplex.html |
| Affymetrix array | www.affymetrix.com/index.affx |
| Illumina array | www.illumina.com/ |
| Haploview software | www.broad.mit.edu/mpg/haploview |

## 犬类基因组变异

通过鉴别拳师狗装配中的 SNP 并与标准狮子狗的 1.5×局部序列相比较已经产生了一个包含 250 万个 SNP 的广泛 SNP 图谱。此外，从 9 只不同品种的狗中每只取出了 100 000 个序列阅读，并将此序列与拳师狗装配序列进行了比较。此 SNP 图谱平均每 1kb 包含 1 个 SNP。此图谱现已向公众开放（http：//www. broad. mit. edu/ftp/pub/papers/dog _ genome/snps _ can-snps _ canfam2 和 http：//www. broad. mit. edu. /mammals/dog)(Lindblad-Toh et al. 2005)。此 SNP 有 98%的平均验证率，比通过比较狮子狗和拳狮狗序列发现的 SNP 比例稍低。它在品种间的通用性大约为 70%，本质上在品种间是随机分布的。

## 单体型结构和关联作图策略

犬科种群历史包含了两个种群瓶颈，比较近的一个瓶颈在过去几百年中导致了品种的产生。这种已确立的事实长期以来使遗传学家推测狗中的连锁不平衡（LD）会很长，因此可用于基因组范围内的关联作图。这已在两项研究中得到证明。最初，Sutter 和同事检查了 5 个具有不同历史的品种中的 LD 范围并报道平均 LD 长度大约为 2Mb（Sutter et al. 2004)。这比人类基因组中典型的 LD 要长 40～100 倍。作为基因组测序工程的一部分，10 个独立的狗品种内品种间的单体型数目和长度作为一个整体也分成 10 个随机的 15Mb 区域而且得以仔细地检查。品种内的单体型和长距离的 LD 和典型的3～6 个品种特异性单体型也被呈现在每个位置上。在整个狗种群中，单体型和 LD 非常短。在每个位置上仅仅发现了 3～5 个祖先单体型，一些单体型甚至在关系很远的品种间也相同。这表明一些致病突变可能是祖传的，不仅存在于密切相关的品种间，还存在于不同的品种间。

然而人类连锁研究需要大约 500 000 间隔相等的 SNP（Kruglyak 1999；HapMap 协会 2003），在狗中 LD 的距离扩大了约 50 倍说明狗的关联研究需要 10 000 个间隔相等的 SNP。为了对其进行估计，人们模拟了一组种群，将其进行校准从而使其序列的 SNP 率和 LD 曲线与真实数据相同，得到了 1 Mb 的序列（Lindblad-Toh 2005）。不同的标记密度是通过随机选择适量的均匀间隔的 SNP 而得以确定的。具有较小等位基因频率（MAF＜20％）的单个 SNP 被选作“疾病基因”并且通过不同标记密度的关联分析对它们作图的能力进行了检测。使用 15 000 个或者 30 000 个基因组范围内的 SNP 分析也得到了相同的结果。在 7500 个 SNP 中观察到了稍低的强度（图 31-3A）。

每项研究所需的狗数也得到了详细的检查（Lindblad-Toh et al. 2005）。对于导致高外显率并且没有表型模拟的简单孟德尔性状的疾病等位基因，使用 20 个假装的事件和 20 个自然对照足以作出一个隐性座位的图谱。对显性性状而言，需要 50 个患病的和 50 个健康的狗。对于多基因性状，检测疾病基因的能力依赖于多种因素，包括等位基因带来的相对风险、等位基因频率和与其他基因之间的相互作用。一个等位基因通过 2～5 倍的倍增因子 $\lambda$ 增加风险的简单模型也已得到了研究（Lindblad-Toh et al. 2005）。在以上的 SNP 密度和显著性阈值条件下，本来使用 100 个患病的和 100 个健康的狗样检测一个座位的能力在 $\lambda=5$ 时是 98％，$\lambda=2$ 时是 50％。事实上，假如一个能增加两倍风险的等位基因在种群中频率适中（＜20％），使用大约 500 只患病的狗和 500 只健康的狗应该足以作出一个座位的图谱（图 31-3B）。

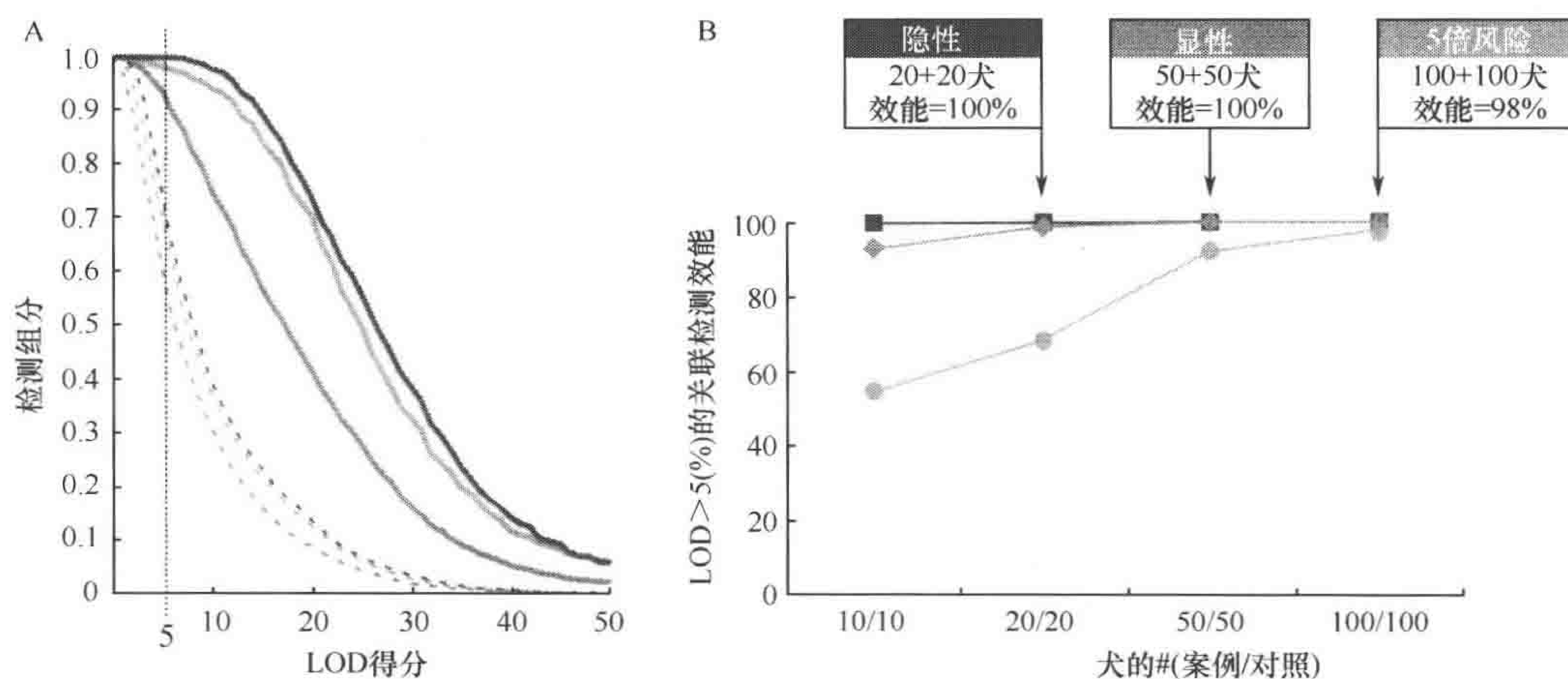



图 31-3 基因组关联作图检测疾病座位的能力。A. 表示多 SNP 单元型（实线），其 SNP 密度等价于 7500（红）、15 000（绿）和 30 000（蓝）个 SNP 的基因组图谱，比单个 SNP（虚线）更易于检测关联性。一个 SNP 代表一个能增加五倍风险的等位基因模型。周围 1Mb 染色体区域的 SNP 基因型由与观察到的品种内变异相对应的 coalescent 模型来模拟。然后根据座位的基因型和遗传模式产生了 100 假装的狗和 100 只自然狗在此染色质区域的二倍体基因型。进行关联分析以检测致病基因是否存在（Lindblad-Toh et al. 2005）。(B) 表示一个 SNP 代表以下三种遗传模式的疾病基因之一：简单孟德尔隐性（蓝）、简单孟德尔显性（红）和五倍风险增加（绿）(见图版)

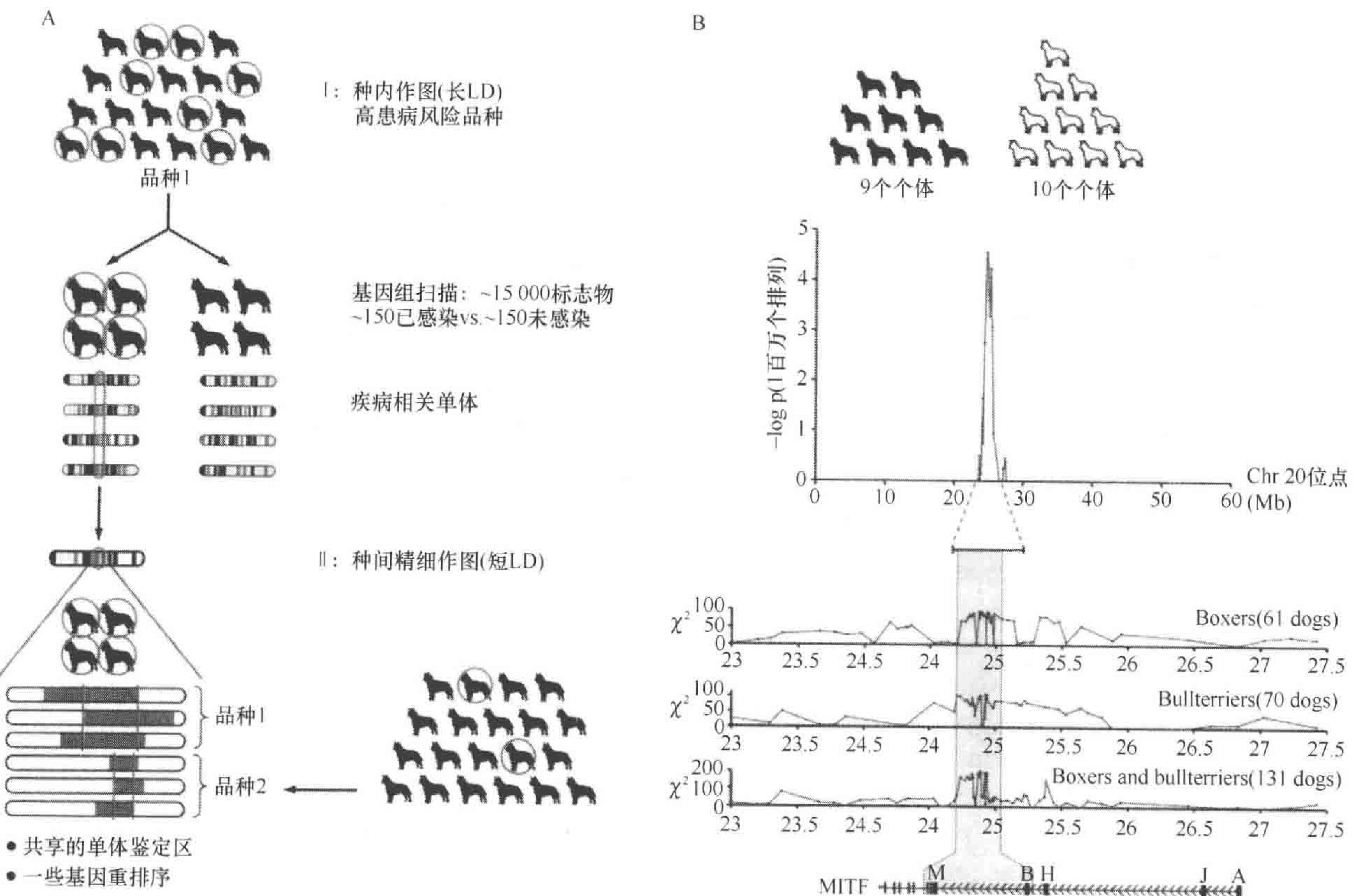



图31-4 两步作图策略鉴定小于0.5Mb的区域。A.表示一个两步作图策略。第一步，在一个患某种疾病风险较高的品种中使用10 000~20 000个SNP对全基因组进行关联作图。第二步，结合一些带有共同祖先的疾病单体型的相关品种在初始品种中对疾病相关区域进行精细作图。这样就能迅速地缩小区域以筛选出一个或多个基因上的部分突变。B.表示对拳师狗中的白领座位进行作图时，在基因组关联作图步骤中采用了10只白领狗和9只纯色狗，从而鉴定出位于20号染色体上的一个约1Mb区域。进行精细作图时使用了另一个品种，从而将相关区域缩短到了100kb。此区域仅包含了*MIFT*的4个外显子

## 两步作图策略（two-tiered mapping）

上面的数据表明只要最初扫描基因组时使用了单个品种并且表型精确，就有可能用少量的样本作出孟德尔性状和复杂性状的图谱。尽管如此，由于存在一些短的祖先单体型片段，所以可以使用两步作图策略，它使用具有相同表型的多个品种进行精细作图，可以将致病突变定位在非常离散的区域上（图 31-4A）。使用两步作图策略的一个非常优秀的例子就是作出了一个导致进行性锥杆细胞退化的基因的图谱。通过传统的微卫星基因组扫描的方法最初将这种疾病绘制在了犬科 9 号染色体上（Acland et al. 1998），然后又用品种组合进行精细作图将此疾病基因定位在了 106 kb 的区域上（Goldstein et al. 2006），此处有一个导致犬科疾病的错义突变。同样的错义突变还存在常染色体隐性形式，这是在一位孟加拉国患者身上发现的（Zangerl et al. 2006）。

## 犬类基因组关联作图工具

为了能用已预测的大约 15 000 个 SNP 进行基因组关联作图，哈佛 Broad 研究所和麻省理工学院与 Affymetrix 公司合作制造了一个基因组 SNP 阵列来测定基因型。最终的质量受控（quality-controlled）（QCed）阵列包含大约 26 000 个有功能的 SNP，这覆盖了整个基因组并且代表了狗种群的多样性。第二代阵列有望包含将近 50 000 个有功能的 SNP。此外，一个包含 26 000 个 SNP 的 Illumina 阵列也正处于制造之中，预测在 2007 年初会投入使用。这些阵列为犬科动物研究团队提供了资源。

最早的 Affymetrix 阵列包含大约 65 000 个 SNP，这覆盖了整个基因组并且代表了品种的多样性。预计大约 70%的 SNP 在任意特定的品种中都具有多态性（Lindblad-Toh et al. 2005），根据这项技术固有的 25%～55%的成功率，最终会有一套15 000～30 000 的有功能的 SNP 通过最初的 QC。阵列原理与人类 500K Affymetrix SNP 阵列相似（Matsuzaki et al. 2004），DNA 被消化掉并且大约 10%的基因组 DNA 大小会在一定范围内与寡聚核苷酸阵列杂交。低复杂度的 DNA 用于减少交叉杂交。（起始 DNA 要具有高分子量以免降解了的 DNA 大小在所选范围之内，这很重要。因此我们建议在进行杂交时使用由血中提取的 DNA 而不是擦拭面颊制备的 DNA。）为了进一步降低噪声，要用多个随机的 25-mer 片段来代表等位基因 A 和 B。一套高效率的 SNP 是根据从多个品种的几百只狗中得到的一套数据而选得。

所选的大约 26 000 个 SNP 的基因组覆盖度非常好。遍及狗的 38 个常染色体的 1-Mb bins 中有 97%包含 5 个以上的 SNP，所有 1-Mb bins 中都至少包含两个 SNP。染色体 X 的覆盖度较小，仅仅 42%的 1-Mb bins 中具有 5 个以上的 SNP，88%的 1-Mb bins 包含至少一个 SNP（Hillbertz et al. 2007；Karlsson et al. 2007），这可能是由于它的遗传多样性较低且重复含量高。平均每只狗中 93%的 SNP 精确度＞99.5%。品种内的多态性 SNP＞70%（MAF＞5%），与预期一致。在品种间，如果使用 4 配子规则和 5%的 MAF，平均单体型块大小估计为 600 kb。尽管如此，SNP 数目更多的下一代阵列会修正这个数目。在每个块内观察到有 3.8 个单体型频率＞5%（图 31-5）。

使用全部约 26 000 个 SNP 可以很容易地将狗按品种进行聚类，每个品种只需要分

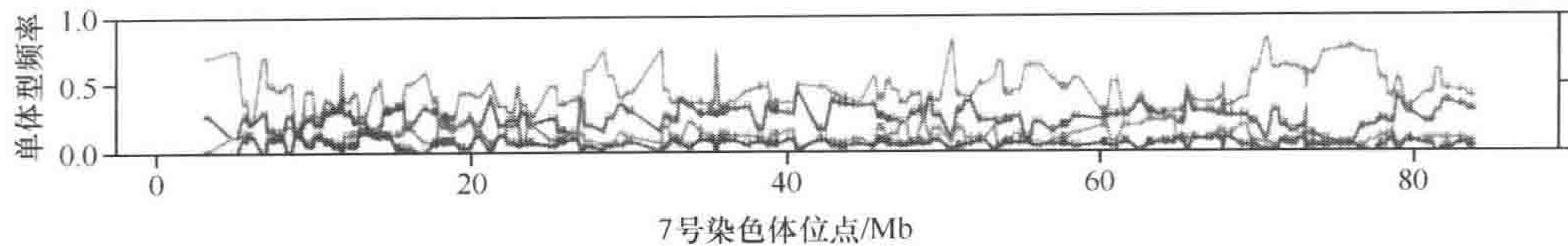



图 31-5 rottweiler 中 7 号染色体上的单体型频率和大小。单体型大小平均为 600 kb，典型的座位有 3～4 个单体型，绝大多数频率为 40%～80%。红色标记为最常见的单体型，蓝色、绿色和紫色为其他单体型（见图版）

析 10 只狗。使用 Eigenstrat 得到的平均漂移时间是 0.3～0.4（Price et al. 2006)。平均漂流时间与 FST（F-统计学，用于衡量种群间的遗传变异）类似，但衡量的是整个基因组。使用这种方法要求对种群内的主要组分进行分析并且种群内具有明确不同的模式世系，这是此法的基础。目前根据 15 个组分已经分出了 16 个品种（Price et al. 2006)。这些品种分类清晰，与以前的数据（Parker et al. 2004）是一致的。相比之下，比较美国和荷兰金毛猎犬得出的漂移时间仅仅是 0.11，与比较高加索人和亚洲人种群时得出的数据接近（HapMap 协会 2003)。这说明狗品种是清楚的实体，与人类种群相似，甚至在品种内也由于生理位置或者领袖者效应而存在强烈的等级。因此配对来自一个种群或者家系的实验组和对照组并进行基因组关联分析时必须要仔细考虑上述情况。

在写这章的时候，阵列已被用于一些单基因性状的作图工作中并提供了一些复杂性状的初级结果。在所有的情况中，使用上述 power 计算中预言的样本数已经在基因组水平上作出了座位的图谱。拳师狗中白领是一种共显性性状，图 4b 显示如何使用 9 个纯色和 10 个白色拳师狗作出它的图谱。鉴定出的区域小于 1 Mb 并且含有 *MITF*（小眼相关转录因子）基因，已知这种基因能在人和老鼠中导致色素沉着紊乱（Udono et al. 2000；Steingrimsson et al. 2004；Karlsson et al. 2007)。

## 多品种间的精细作图有助于精确地鉴定突变

大多数基因组关联作图项目只能鉴定出大约 1 Mb 的区域，通常只包括少数基因，而多品种间的精细作图则可以提供更精确的信息。鉴定拳师狗中的白领基因时，在精细作图步骤中使用了另一种产生此种性状的品种，找到了一段大约 100 kb 不重组的片段，它含有 4 个外显子、几个保守的非编码元件和疾病突变（图 31-4B）（Karlsson et al. 2007)。这种共享的单体型与先前描述的 prcd 研究中观察到的非常相似，那项研究中也使用了多个品种（Goldstein et al. 2006)。

看起来两步作图策略用于复杂性状时很有效，assuming 表型也可以有效地得到确定以显示共享的风险因子。自然条件下，越是密切相关的品种越可能共享相同的风险等位基因，但如果共享的单体型在关系较远的品种中也频繁出现（Lindblad-Toh et al. 2005；Hillbertz et al. 2007)，就要考虑远缘品种是否也有同样的表型。

### 复杂性状的变异基础

在人类和狗中很少发现真正具有复杂性状的变体。因此，可以假设一些此类突变本质上是受到调控的。尤其是在狗中最为可能，狗受到了强烈的人工选择，饲养者不大可能容忍由于编码区突变而在狗中产生的有害突变。

从产生简单遗传的犬科疾病的变体类型中得到的信息非常重要。一些表型与犬科特异性的短散布核元件（SINE）有关（Minnick et al. 1992；Bentolila et al. 1999；Vassetzky and Kramerov et al. 2002）。这些反转座子产生于一种 tRNA-Lys，在整个犬科基因组中都频繁出现（Coltman and Wright 1994；Bentolila et al. 1999；Kirkness et al. 2003）。除了遗传性发作性睡眠症外，SINEC-CF 元件的异常插入还和拉布拉多犬的中央核肌病（Pele et al. 2005）及一些品种中的灰或"鸫类"领有关。

新型变体相关疾病的另一个有趣的例子是一种癫痫症，它与人类肌阵挛性癫痫疾病相似。Lohi 和合作者研究了小型钢毛腊肠狗，指出这种病是由 *Epm2b*（*Nhlrc1*）基因中一种不稳定的 dodecamer 重复扩展导致的（Lohi et al. 2005）。虽然已有报道说三核苷酸重复扩展与一些人类神经紊乱症有关（Kazemi-Esfarjani et al. 1995），但这是第一个关于重复扩展能在人类之外的其他物种中导致疾病的报道。

这些突变可能在狗的复杂性状中也起着作用（Kirkness 2006）。这样对狗的性状进行作图就变得简单了，但寻找潜在的种系变体则不简单。依靠可以使用的多个品种来高度准确地鉴定含有突变的区域将有助于鉴定许多变体，这些变体将用于检测相关基因是否具有功能。

### 处于选择作用下的区域的鉴定

一些性状在必需的品种标准影响之下被强烈地选择下来。其中一些是形态建成的或者其他的生理性状，也包括行为性状。可以假设这些选择会在基因上留下足迹，因为特定的等位基因被固定了下来或者过量存在。分析来自一个品种的多个体的基因组阵列数据可以鉴定出处于选择作用之下的基因组区域。尽管如此，当前存在的证据暗示这些座位经常少于 1 Mb（Hillbertz et al. 2007；Karlsson et al. 2007）。这与单纯通过遗传漂变而纯合起来的随机单体型大小相同（Karlsson et al. 2007）。因而，鉴定这些区域要求仔细地分析以除去统计学噪声信号。看起来最有前途的方法是在具有相同选择表型的多个远缘品种中比较这些纯合区域。

## 结论

狗类系统正处在一个十字路口。研究团体希望得到的遗传工具和资源现在都已经可以利用了。在犬科系统中，人们使用图谱、序列、标记和芯片来鉴定在疾病的易感性、发展和结果过程中发挥着重要作用的基因。现在临床医生和遗传学家的任务是确定和搜集家庭和种群资源使这些进展产生最大的作用。现在遍及全世界的实验室都在进行这些工作。由于犬科遗传学研究团体的工作，我们得到信息的质量和数量都达到了一个新的

水平。我们现在的任务就是有效地使用这些资源，合理选择性状和疾病进行作图，这样就会不断地扩展关于我们自身的知识。狗是我们最亲密的同伴，哺乳动物生物学是一个整体。

## 致谢

我们感谢狗的主人、饲养者和支持者们，他们不断地向我们提供宠物样本和信息。本工作由国家人类基因组内部项目、国家人类基因组研究所拨款，资助编号为HG03067 和 HG003969，同时受到了美国康奈尔犬科健康基金的支持。

## 参考文献

Acland G.M., Blanton S.H., Hershfield B., and Aguiree G.D. 1994. XLPRA: A canine retinal degeneration inherited as an X-linked trait. *Am. J. Med. Genet.* **52:** 27–33.

Acland G.M., Ray K., Mellersh C.S., Gu W., Langston A.A., Rine J., Ostrander E.A., and Aguirre G.D. 1998. Linkage analysis and comparative mapping of canine progressive rod-cone degeneration (*prcd*) establishes potential locus homology with retinitis pigmentosa (RP17) in humans. *Proc. Natl. Acad. Sci.* **95:** 3048–3053.

——. 1999. A novel retinal degeneration locus identified by linkage and comparative mapping of canine early retinal degeneration. *Genomics* **59:** 134–142.

American Kennel Club. 1998. *The complete dog book*, 19th edition revised. Howell Book House, New York, p. 790.

American Veterinary Medical Association. 2002. *U.S. pet ownership and demographics sourcebook.* American Veterinary Medical Association, Schaumburg, Illinois.

Bentolila S., Bach J.M., Kessler J.L., Bordelais I., Cruaud C., Weissenbach J., and Panthier J.J. 1999. Analysis of major repetitive DNA sequences in the dog (*Canis familiaris*) genome. *Mamm. Genome* **10:** 699–705.

Cavelli-Sforza L.L., Menozzi P., and Piazza A. 1994. *The history and geography of human genes.* Princeton University Press, Princeton, New Jersey.

Clark L.A., Wahl J.M., Rees C.A., and Murphy K.E. 2006. Retrotransposon insertion in *SILV* is responsible for merle patterning of the domestic dog. *Proc. Natl. Acad. Sci.* **103:** 1376–1381.

Coltman D.W. and Wright J.M. 1994. *Can* SINEs: A family of tRNA-derived retroposons specific to the superfamily Canoidea. *Nucleic Acids Res.* **22:** 2726–2730.

Goldstein O., Zangerl B., Pearce-Kelling S., Sidjanin D.J., Kijas J.W., Felix J., Acland G.M., and Aguirre G.D. 2006. Linkage disequilibrium mapping in domestic dog breeds narrows the progressive rod–cone degeneration interval and identifies ancestral disease-transmitting chromosome. *Genomics* **18:** 541–550.

Hillbertz S.H., Isaksson M., Karlsson E.K., Hellmén E., Rosengren Pielberg G., Savolainen P., Wade C.M., von Euler H., Gustafson U., Hedhammar A., et al. 2007. A duplication of FGF3, FGF4, FGF9 and ORAOV1 causes the hair ridge and predisposes to dermoid sinus in Ridgeback dogs. *Nat. Genet.* (in press).

International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Jonasdottir T.J., Mellersh C.S., Moe L., Heggebo R., Gamlem H., Ostrander E.A., and Lingaas F. 2000. Genetic mapping of a naturally occurring hereditary renal cancer syndrome in dogs. *Proc. Natl. Acad. Sci.* **97:** 4132–4137.

Karlsson E.K., Baranowska I., Wade C.M., Nicolette H.C., Hillbertz S., Zody M.C., Anderson N., Biagi T.M., Patterson N., Rosengren Pielberg G., et al. 2007. Genome-wide association mapping in dogs—A powerful approach for gene discovery. *Nat. Genet.* (in press).

Kazemi-Esfarjani P., Trifiro M.A., and Pinsky L. 1995. Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor: Possible pathogenetic relevance for the $(CAG)_n$-expanded neuronopathies. *Hum. Mol. Genet.* **4:** 523–527.

Kirkness E.F. 2006. SINEs of canine genomic diversity. In *The dog and its genome* (ed. E.A. Ostrander et al.), pp. 209–219. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Kirkness E.F., Bafna V., Halpern A.L., Levy S., Remington K., Rusch D.B., Delcher A.L., Pop M., Wang W., Fraser C.M., and Venter J.C. 2003. The dog genome: Survey sequencing and comparative analysis. *Science* **301:** 1898–1903.

Koskinen M.T. 2003. Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. *Anim. Genet.* **34:** 297–301.

Koskinen M.T. and Bredbacka P. 2000. Assessment of the population structure of five Finnish dog breeds with microsatellites. *Anim. Genet.* **31:** 310–317.

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22:** 139–144.

Lin L., Faraco J., Li R., Kadotani H., Rogers W., Lin X., Qiu X., de Jong P.J., Nishino S., and Mignot E. 1999. The sleep disorder canine narcolepsy is caused by a mutation in the *hypocretin* (*orexin*) receptor 2 gene. *Cell* **98:** 365–376.

Lindblad-Toh K., Wade C.M., Mikkelsen T.S., Karlsson E.K., Jaffe D.B., Kamal M., Clamp M., Chang J.L., Kulbokas E.J., III, Zody M.C., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.

Lohi H., Young E.J., Fitzmaurice S.N., Rusbridge C., Chan E.M., Vervoort M., Turnbull J., Zhao X.C., Ianzano L., Paterson A.D., et al. 2005. Expanded repeat in canine epilepsy. *Science* **307:** 81.

Matsuzaki H., Dong S., Loi H., Di X., Liu G., Hubbell E., Law J., Berntsen T., Chadha M., Hui H., et al. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1:** 109–111.

Mignot E., Wang C., Rattazzi C., Gaiser C., Lovett M., Guilleminault C., Dement W.C., and Grumet F.C. 1991. Genetic linkage of autosomal recessive canine narcolepsy with a mu immunoglobulin heavy-chain switch-like segment. *Proc. Natl. Acad. Sci.* **88:** 3475–3478.

Minnick M.F., Stillwell L.C., Heineman J.M., and Stiegler G.L.

1992. A highly repetitive DNA sequence possibly unique to canids. *Gene* **110:** 235–238.
O'Rourke K. 2005. Mining the canine genome. Identification of genes helps breeders and researchers. *J. Am. Vet. Med. Assoc.* **226:** 863–864.
Ostrander E.A. and Friedrichsen D.M. 2004. Genetic factors: Finding cancer susceptibility genes. In *Clinical oncology*, 3rd edition (ed. M.D. Abeloff et al.), pp. 253–267. Elsevier Churchill Livingstone, Philadelphia, Pennsylvania.
Ostrander E.A. and Kruglyak L. 2000. Unleashing the canine genome. *Genome Res.* **10:** 1271–1274.
Ostrander E.A. and Wayne R.K. 2005. The canine genome. *Genome Res.* **15:** 1706–1716.
Ostrander E.A., Markianos K., and Stanford J.L. 2004. Finding prostate cancer susceptibility genes. *Annu. Rev. Genomics Hum. Genet.* **5:** 151–175.
Parker H.G. and Ostrander E.A. 2005. Canine genomics and genetics: Running with the pack. *PLoS Genet.* **1:** e58.
Parker H.G., Kim L.V., Sutter N.B., Carlson S., Lorentzen T.D., Malek T.B., Johnson G.S., DeFrance H.B., Ostrander E.A., and Kruglyak L. 2004. Genetic structure of the purebred domestic dog. *Science* **304:** 1160–1164.
Patterson D. 2000. Companion animal medicine in the age of medical genetics. *J. Vet. Intern. Med.* **14:** 1–9.
Patterson D.F., Haskins M.E., and Jezyk P.F. 1982. Models of human genetic disease in domestic animals. *Adv. Hum. Genet.* **12:** 263–339.
Patterson D.F., Haskins M.E., Jezyk P.F., Giger U., Meyers-Wallen V.N., Aguirre G., Fyfe J.C., and Wolfe J.H. 1988. Research on genetic diseases: Reciprocal benefits to animals and man. *J. Am. Vet. Med. Assoc.* **193:** 1131–1144.
Pelé M., Tiret L., Kessler J.L., Blot S., and Panthier J.J. 2005. SINE exonic insertion in the *PTPLA* gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Hum. Mol. Genet.* **14:** 1417–1427.
Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A., and Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38:** 904–909.
Rosenberg N.A., Pritchard J.K., Weber J.L., Cann H.M., Kidd K.K., Zhivotovsky L.A., and Feldman M.W. 2002. Genetic structure of human populations. *Science* **298:** 2381–2385.
Sargan D.R. 2004. IDID: Inherited diseases in dogs: Web-based information for canine inherited disease genetics. *Mamm. Genome* **15:** 503–506.
Sargan D., Aguirre-Hernandez J., Galibert F., and Ostrander E.A. 2007. An extended microsatellite set for linkage mapping in the domestic dog. *J. Hered.* (in press).
Savolainen P., Zhang Y.P., Luo J., Lundeberg J., and Leitner T. 2002. Genetic evidence for an East Asian origin of domestic dogs. *Science* **298:** 1610–1613.
Sidjanin D.J., Lowe J.K., McElwee J.L., Milne B.S., Phippen T.M., Sargan D.R., Aguirre G.D., Acland G.M., and Ostrander E.A. 2002. Canine CNGB3 mutations establish cone degeneration as orthologous to the human achromatopsia locus ACHM3. *Hum. Mol. Genet.* **11:** 1823–1833.
Steingrimsson E., Copeland N.G., and Jenkins N.A. 2004. Melanocytes and the microphthalmia transcription factor network. *Annu. Rev. Genet.* **38:** 365–411.
Sutter N.B., Eberle M.A., Parker H.G., Pullar B.J., Kirkness E.F., Kruglyak L., and Ostrander E.A. 2004. Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res.* **14:** 2388–2396.
Udono T., Yasumoto K., Takeda K., Amae S., Watanabe K., Saito H., Fuse N., Tachibana M., Takahashi K., Tamai M., and Shibahara S. 2000. Structural organization of the human microphthalmia-associated transcription factor gene containing four alternative promoters. *Biochim. Biophys. Acta* **1491:** 205–219.
Vassetzky N.S. and Kramerov D.A. 2002. CAN—A pan-carnivore SINE family. *Mamm. Genome* **13:** 50–57.
Vilà C., Savolainen P., Maldonado J.E., Amorim I.R., Rice J.E., Honeycutt R.L., Crandall K.A., Lundeberg J., and Wayne R.K. 1997. Multiple and ancient origins of the domestic dog (see comments). *Science* **276:** 1687–1689.
Wade C., Karlsson E.K., Mikkelsen T.S., Zody M.C., and Lindblad-Toh K. 2006. The dog genome: Sequence, evolution and haplotype structure. In *The dog and its genome* (ed. E.A. Ostrander et al.), pp. 179–207. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
Wayne R.K., Nash W.G., and O'Brien, S. J. 1987a. Chromosomal evolution of the Canidae. I. Species with high diploid numbers. *Cytogenet. Cell Genet.* **44:** 123–133.
———. 1987b. Chromosomal evolution of the Canidae. II. Divergence from the primitive carnivore karyotype. *Cytogenet. Cell Genet.* **44:** 134–141.
Wayne R.K., Geffen E., Girman D.J., Koeppfli K.P., Lau L.M., and Marshall C.R. 1997. Molecular systematics of the Canidae. *Syst. Biol.* **46:** 622–653.
Yuzbasiyan-Gurkan V., Blanton S.H., Cao V., Ferguson P., Li J., Venta P.J., and Brewer G.J. 1997. Linkage of a microsatellite marker to the canine copper toxicosis locus in Bedlington terriers. *Am. J. Vet. Res.* **58:** 23–27.
Zangerl B., Goldstein O., Philp A.R., Lindauer S.J., Pearce-Kelling S.E., Mullins R.F., Graphodatsky A.S., Ripoll D., Felix J.S., Stone E.M., et al. 2006. Identical mutation in a novel retinal gene causes progressive rod–cone degeneration in dogs and retinitis pigmentosa in humans. *Genomics* **26:** 551–563.

# 互联网资源

**http://www.genome.gov/125134** ENCODE, Project Background, National Human Genome Research Institute.

# 32 黑 猩 猩

Tarjei S. Mikkelsen, Michael C. Zody, and Kerstin Lindblad-Toh

*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142*



## 简介

作为与我们进化关系最近的现存生物，黑猩猩（Chimpanzee，*Pan troglodytes*）为各个人种及其历史提供了一个独特的视角。我们所特有的可遗传生物学性状，如独特的骨骼结构、认知能力和一些疾病敏感性，最终都是由人类和黑猩猩基因的不同造成的。比较与分析我们的基因组序列有助于解释这些不同和产生它们的突变过程及选择压力。此外，作为与现代人种群最接近的外群，黑猩猩在人类遗传变异研究中有着特别的作用。

一个以美国为基础的协会（黑猩猩测序和分析协会 2005）使用相当于普通黑猩猩（*Pan troglodytes*）基因组长度 4 倍的散弹片段产生了最早的黑猩猩基因组序列。在写本章的同时，进一步的测序工作也已完成，得到了总共 6. 5-fold 的散弹片段和一个基于细菌人工染色体（BAC）的物理图谱（Warren et al. 2006）。其他可以利用的基因组资源也越来越多，包括用来自另外一个个体的 21 号染色体装配而成的 BAC（Watanabe et al. 2004）、两个 Y 染色体序列（Hughes et al. 2005；Kuroki et al. 2006）、从 13 000个已知基因中 PCR 扩增出的外显子（Nielsen et al. 2005）、cDNA 序列（Hellmann et al. 2003）和另外的西非和中非黑猩猩的 light 散弹片段（黑猩猩测序和分析协会 2005）。几个研究团队预先估计的基因和其基因组定位也已能使用（http://genome. ucsc. edu/；http://www. ensembl. org；http://www. ncbi. nih. gov）。表 32-1 总结了可用的基因组资源。

**表 32-1　黑猩猩基因组资源**

| |
|---|
| 基因组序列和注释 |
| http://www.ncbi.nlm.nih.gov/genome/guide/chimp/ |
| http://www.ensembl.org/Pan_troglodytes/ |
| http://genome.ucsc.edu |
| SNP |
| http://www.broad.mit.edu/mammals/chimp/SNP/ |
| 黑猩猩基因组研究概况 |
| http://www.nature.com/nature/focus/chimpgenome/ |

这儿我们描述了人类和黑猩猩基因组比较分析的实际用途和从中得出的观点。除非另作说明，本章的辅助性数据都可以在原稿和出版有最初的猩猩基因组序列的补充材料中找到（黑猩猩测序和分析协会 2005）。

## 基因组序列草图的实际用途

在使用黑猩猩基因组序列进行特定的分析之前，要考虑如何使基因组草图中存在的错误降到最低。这在比较密切相关的物种，如人类和黑猩猩特别重要，相对于二者很小的基因分歧量，这里即使非常小的错误率也会变得非常显著。

序列质量评分为在单核苷酸水平上鉴定或忽略基因组错误提供了一个必需的工具，并且从与基因组序列相同的源上可以得到它。在构建共有序列时，基因组装配软件使用从个体序列中得到的质量评分来衡量不一致的阅读提供的数据，并报告在这个共有序列中每个核苷酸的质量评分。现在使用的质量评分遵循了 PHRED 评分系统（Ewing and Green 1998），它与对应核苷酸的估计精确性呈对数关系（10 时为 90%，20 时为 99%，30 时为 99.9%等）。根据这个系统，人类和黑猩猩中的低分歧率会造成非常高的分数截点。尽管如此，实验中观察到大多数序列错误都出现在紧挨着容易检测的人工假象的位置，这导致质量评分（NQS）过滤器的产生（Altshuler et al. 2000），它接受低质量的核苷酸，只要该苷酸附近没有明显的错误。各种各样的 NQS 标准已经被设计出来用于人-黑猩猩比较。我们发现中央核苷酸的质量评分为 30，两侧各 5 个核苷酸是 25，并且对两侧的替换数没有限制，就足以将倍基因组大小的散弹装配序列从仅仅是“完成”提升到一个非常精确的水平（精确度大于 99.99%）。而且，还能确认任何特定分析的定量结果对这些参数是否支持。

用于测定黑猩猩基因组序列的 WGS 法不能将供体中的两种单体型分开，这将导致质量评分系统产生偏差。变种等位基因导致重叠的阅读发生冲突，在某些情况下很难用基因组装配算法来区分是测序错误还是真正的多态性位点。结果，相对于测序个体中的纯合位点，现在装配成的黑猩猩的基因组杂合位点的质量评分偏低。相反，人类参考序列是用鸟枪法测定大的、单个单体型克隆得到的（国际人类基因组测序协会 2004）。虽然这种偏向对大多数比较分析的影响可以忽略，但如果所用方法对家系特异性的分歧率中的小变异敏感就必须考虑这种影响（见下文）。

错误的序列重叠是装配草图中另一种常见的人工假象，或者更精确地说就是如何辨

别它们。WGS装配算法通过合并重叠的read产生了紧邻的序列（contig）。不完整的覆盖范围导致两个紧邻序列间产生缺口，丢失序列的长度通常可由跨过缺口的末端配对的read来估计。在一些情况下，两个紧邻序列包含冗余区域，这些区域由于序列不一致而不能得以合并，这些不一致序列通常是由于低质量的阅读产生的。这种情况通常由软件识别并被标记为负缺口长度来估计，但是当每个染色体的紧邻序列被合为一个连续序列的时候，这通常被用于装配后分析和表示，负链缺口就会被转换为100个碱基的正链缺口（用通用碱基N来表示）。如果将这种连续的序列与人类基因组进行比对，缺口两侧的冗余区域就会表现为黑猩猩序列中的人工“插入”。现在已经可以利用从代表原始WGS紧邻序列的连续序列作出的图谱装配任意草图，在推断大范围插入/缺失片段（indel）时是一个避免这种假象的有效工具。

在一些情况下，研究者要将个体黑猩猩序列阅读和人类基因组进行比对，而不是从已装配了的基因组序列开始（Hellmann et al. 2005；Patterson et al. 2006）。read通常用BLASTZ（Schwartz et al. 2003）和ARACHNE（Jaffe et al. 2003）等软件来比对，仅仅不能正确确定位置的高质量分数的read才被保留下来作进一步的分析。这种方法优点是既能用于计算多个物种或种群的分歧又能分析其多样性比率，而不会从潜在的不同装配算法中引入未知偏向。read比对也已被用于提示人类基因组和黑猩猩基因组中的大范围重排，这些重排是由于大的插入使克隆的末端配对序列位置紊乱造成的。

## 人类和黑猩猩的遗传分歧

由于遗传人类和黑猩猩中大多数核苷酸都是相同的，发现的不同几乎总是代表一个单突变。通过序列比对来推断人类和黑猩猩基因组的不同可以发现几乎所有的遗传改变，这些改变发生在人类和黑猩猩家系的进化过程中。

这些差异中最为常见的是单核苷酸替换。总体上，共有1.23%的同源核苷酸不同，但是这些替换在基因组中分布并不均匀，这很大部分是由于环境依赖性和突变率的区域差异。例如，虽然CpG二核苷酸在基因组中仅占2%，它们却产生了25%的核苷酸替换。在一个大的尺度上，分歧率在1Mb的片段上从0.99%波动到1.54%（25th～75th四分位数范围；图32-1）。端粒中10Mb的同源序列积累替换平均比基因组其他部分多了15%，潜在地反映了重组介导的突变（Hellmann et al. 2005）。

核苷酸插入和缺失比替代少，但是它们总体上影响的序列更多。在人和黑猩猩基因组中共有500万～600万个缺失插入突变，导致的常染色质基因组中产生了40～45Mb的差异。大多数的插入缺失都很小（98.6%都小于80 bp），但是那些最大的插入缺失很少含有疾病序列（约70 000大于80 bp的插入缺失组成了75%的家系特异性序列）。

可转座元件是在人类和黑猩猩基因组中产生不同种类的插入缺失的原因。最显著的不同是黑猩猩染色体上重复出现了大的近端帽子（Yunis and Parkash 1982）。常染色质序列也显示所有较大种类的可转座元件的插入具有家系特异性。短的灵长类特异性的Alu元件活性在人类中增加了三倍，在人类基因组中能发现超过7000个人类特异性的Alu插入。较长的SVA和LINE-1元件在人类和黑猩猩中以相似的比率插入，每个基

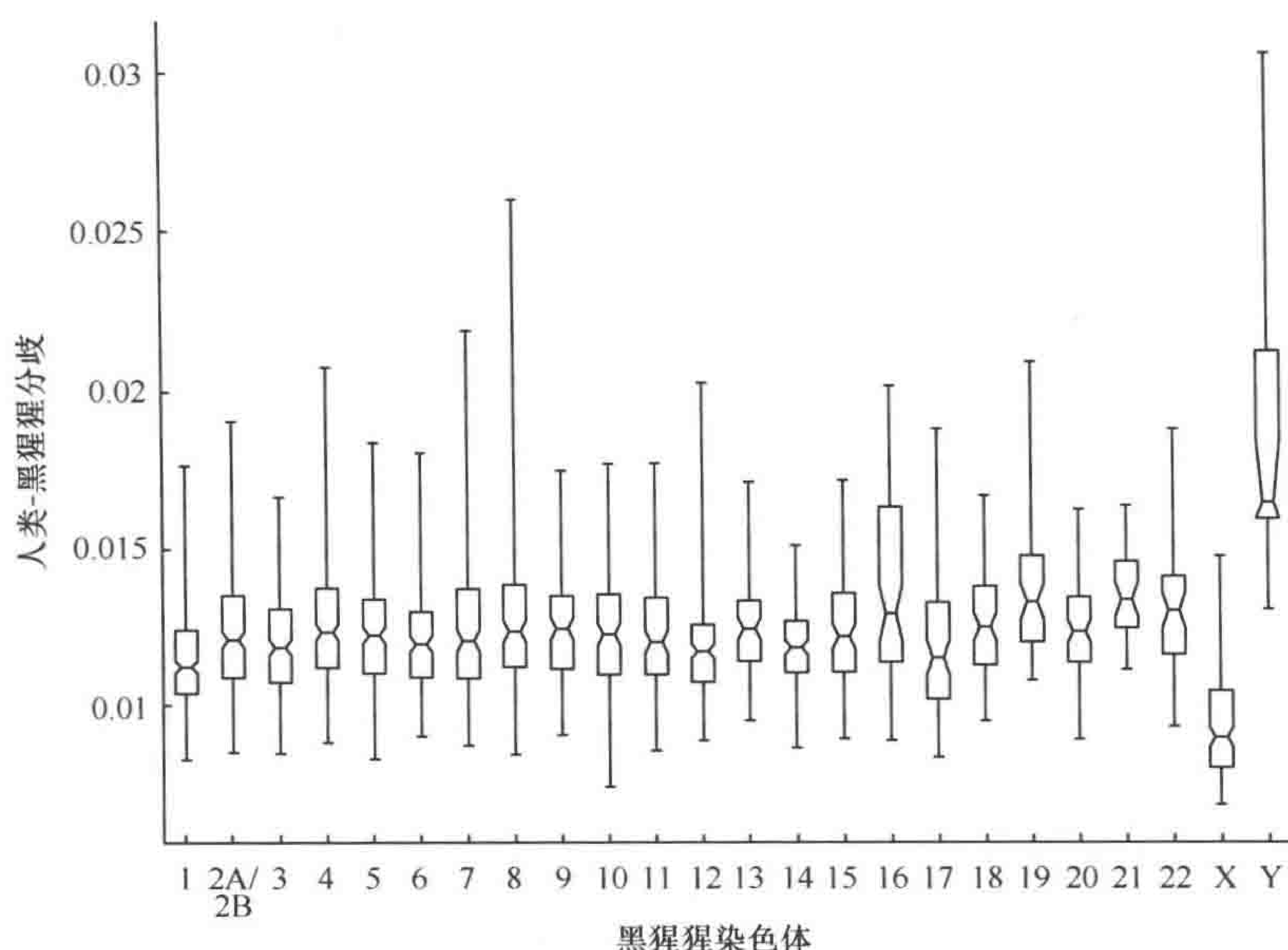



图 32-1　1Mb 片段中人类-黑猩猩分歧的分布，用框来表示。框的边缘表示四分线；刻痕表示平均数的标准误；竖线表示程度（授权复制自黑猩猩测序和分析协会，2005）

因组中分别还有 1000～2000 个新的插入。内源性反转录病毒存在很广泛，但在人类家系中不存在，仅有一种反转录病毒（HERV-K）产生了少于 100 种人类特异性插入。相比之下，黑猩猩基因组携带了几百种插入，这些插入看起来是起源于新型反转录病毒的多个独立的种系（Yohn et al.　2005）。

大范围的染色体重排在两种基因组间形成了最少但却是最戏剧性的不同。早期的研究揭示了人类和黑猩猩的 9 种肠周围的倒置和人类家系中两个祖先染色体的融合（Yunis and Prakash　1982）。使用配对末端序列作图和 WGS 装配来研究结构变异确定了这些重排的定位并提示了其他数百个倒置、片段复制和缺失（Newman et al.　2005）。图 32-2 显示了人类和黑猩猩基因组之间的线性图。

## 现存和祖先的遗传变异对分歧的影响

因为已经测序的人类和黑猩猩染色体在每个核苷酸位置仅仅代表一个现存的等位基因，在人类和黑猩猩之间观察到的不同未必是固定的不同，而可能是人类种群或者黑猩猩种群中的变体等位基因。

两个同源序列间的遗传分歧与产生它们的最后共同祖先的历史成正比，如果这些差异由于新突变的原因而以恒定的速率积累（Zuckerkandl and Pauling　1965）。假设没有选择，在人类和黑猩猩种群中观察到的分歧核苷酸比例为$1-(T_H+T_C)/(2\times T_{HC})$，这里 $T_H$ 是人类种群中一个染色体片段的 TLCA 平均值，$T_C$ 是黑猩猩种群中的 TLCA，$T_{HC}$是人类和黑猩猩的 TLCA，根据聚理论（Rosenberg and feldmann　2002），人类和黑猩猩种群中的预期 TLCA 是 $4\times N_e\times g$，其中 $N_e$ 是有效种群大小（人类中估计为

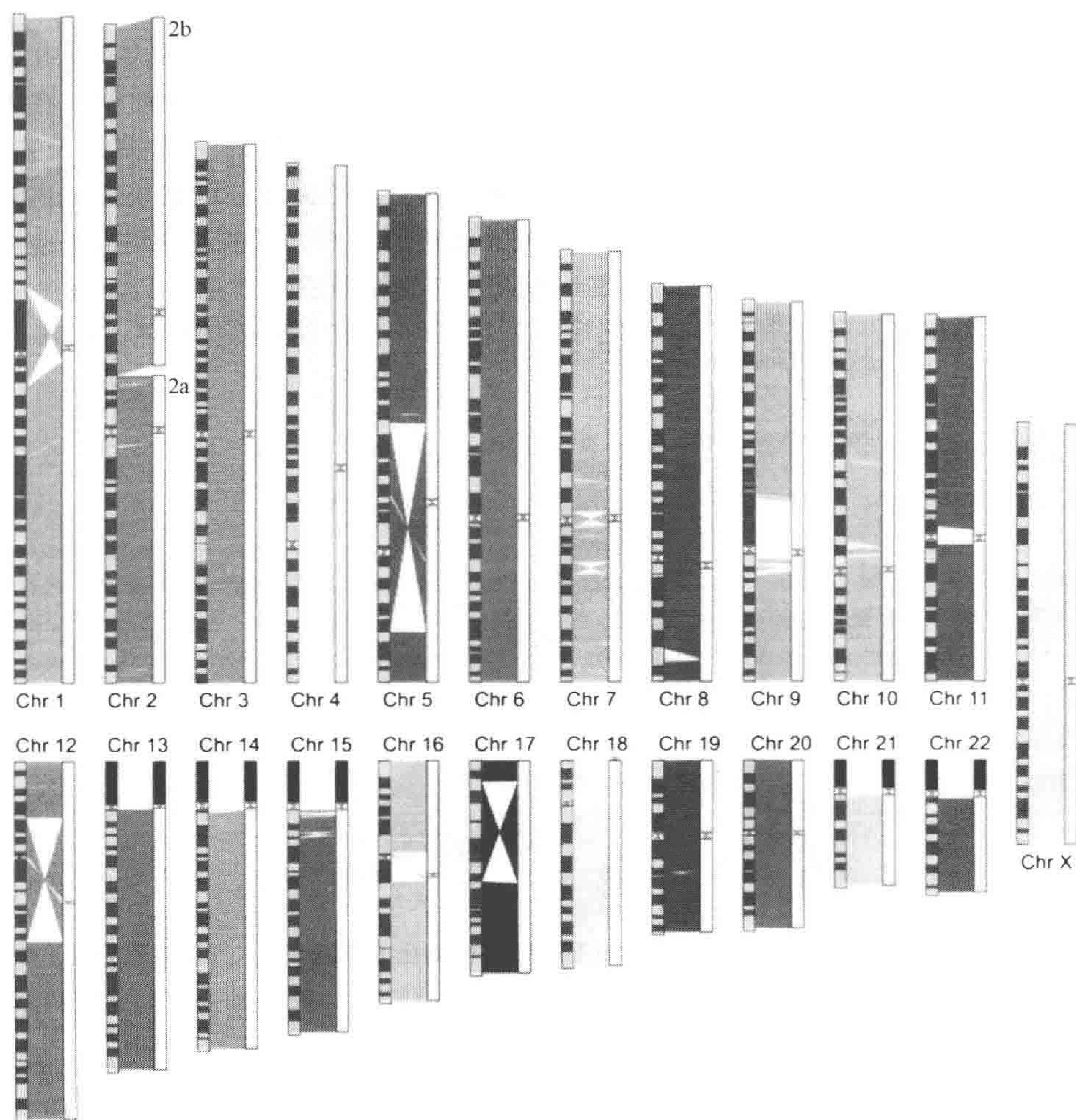



图 32-2 人类和黑猩猩染色体的线性图谱。人类的和黑猩猩染色体呈共线性。对应于两个黑猩猩染色体（2a 和 2b）的人类 2 号染色体和 1 号、4 号、5 号、7 号、9 号、12 号、15 号、16 号、17 号和 18 号染色体上的大倒置片段例外，9 号和 16 号染色体上的倒置主要局限在异染色体质区。线性图谱由 clustering PatternHunter(http://www.bioinformaticssolution.com)alignment 产生，分辨率大于 200 kb。左图表示具有 G 带模式的人类染色体，右图表示黑猩猩的染色体

10 000，黑猩猩为 10 000～20 000)，g 是世代时间（假定为 25 年），$T_H$ =100 万年 (Myr)，Tc=1～2 Myr。假如 $T_{HC}$ =7 Myr(Glazko and Nei 2003)，固定差异的基因组比例为 0.78～0.86。因此在确定特定的替代与人类进化的关系时要测定两个物种中多个个体的基因型，这一点非常重要。

对于其他类型的遗传分歧，现存变异的影响更大。根据结构变异的初始研究，估计相对于人类基因组，黑猩猩基因组中长度大于 12 kb 的缺失，固定下来的差异可能低至 0.67（Newman et al. 2005)。这可能反映了我们基因组中某些区域经常发生的高比率

的结构性重排。

虽然突变率的不同可以解释一部分分歧率区域差异，另一个重要的力量是祖先种群的遗传漂变。对于二倍体生物，同源序列 TLCA 必须大于物种形成的时间，在人类和黑猩猩中估计平均约为 7 百万年（图 32-3）。由于重组，TLCA 并不总是遍及染色体，而是在平均长度小于 10kb（现代非洲种群中连锁不平衡的长度，可能是人类祖先中存在的上限；Reich et al. 2001）的片段之间有着随机的不同。根据观察到的分歧率，估计黑猩猩染色体片段的 TLCA 范围走过 4 百万年。这个范围如此之大以至于一些研究人员提出人类和黑猩猩家系最初分开之后又发生了一个或多个杂交事件（Patterson et al. 2006）。

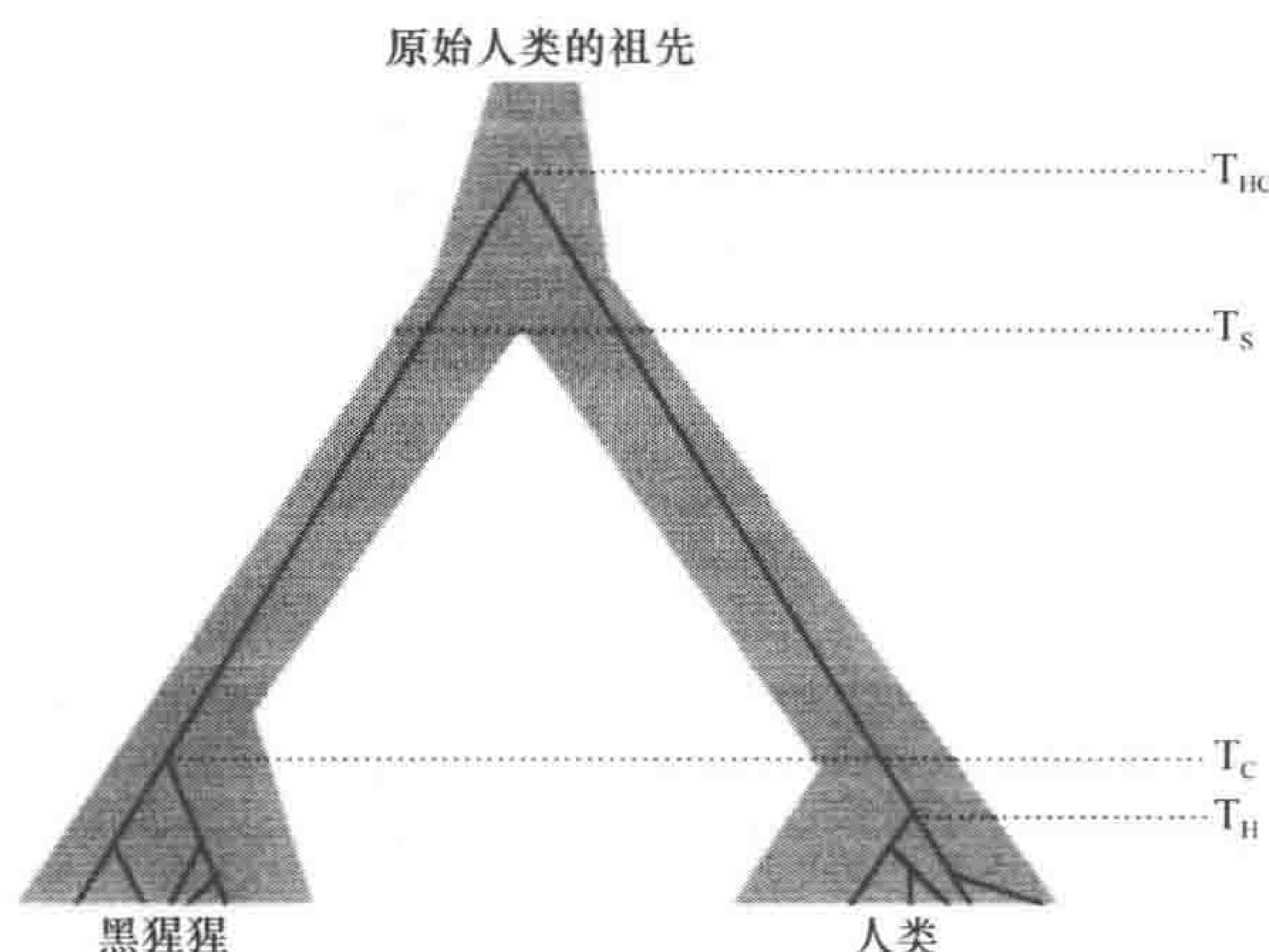



图 32-3 人类和黑猩猩中都存在的单个非重组染色体简化的种系发生。$T_{HC}$表示距离产生片段的人类和黑猩猩的最后共同祖先的时间。$T_H$ 和 $T_C$ 分别表示人类和黑猩猩的 TLCA，已产生片段的每一个现存拷贝都来源于此。$T_S$ 表示距离人和黑猩猩形成的时间

## 祖先等位基因的鉴定

多态性位点的祖先等位基因是所有人最后的共同祖先携带的等位基因。互补的等位基因必须来自于突变推迟了 LCA，被称为衍生性等位基因。

黑猩猩基因组序列是用于等位基因分类的有力工具，特别是人类种群中的 SNP 分类。因为 98.8%的同源核苷酸通过遗传都是相同的，所以可以假设与黑猩猩基因组匹配的人类等位基因是祖先的。依赖于序列比对的严格性，我们使用最初的基因组序列可以确定 dbSNP 中 80%～90%的人类变种是祖先的还是衍生出来的。其他的黑猩猩和密切相关的外围种群，如大猩猩和猩猩的序列数据有助于校订和深化这些任务。

当然，要知道使用黑猩猩基因组确定的祖先等位基因是试探性的。这些任务的错误率可以由同源的黑猩猩核苷酸与人类的一个或两个等位基因而不是人类种群中的祖先等位基因匹配的可能性来估计。这可能发生在如下两种简单情况下：衍生性等位基因固定在了相同的点上，然后发生了一个倒转突变，这个突变仍旧是分离的。涉及更多突变的情况是可能存在的但是可能性至少要低两个数量级。

典型的 SNP 错误率为 0.5%。在人类中等位基因为 CpG 和 TpG 而黑猩猩序列为 TpG 的那些 SNP 例外。这很大一部分是由于一个祖先 CpG 二核苷酸产生两个独立的脱氨基作用造成的，这个 CpG 二核苷酸即是广为人知的突变热点（见上文，人类和黑猩猩的遗传分歧）。在同源黑猩猩序列为 TpG 的人类 CpG 背景中的人类 SNP 占据了 12% 的人类变种，估计错误率为 9.8%。在所有的 SNP 中，平均错误率估计约为 1.6%。

将人类等位基因分为祖先的和衍生的在遗传变异研究中有很多应用，包括人口统计学推断、自然选择和古表型。例如，群体遗传学中一个很好的结果说明，对于随机种内交配的大小不变种群，等位基因是祖先的可能性与它的频率相等（Watterson and Guess 1977）。我们使用 Affymetrix 在 54 个不同个体中测定了大约 120 000 个 SNP 的基因型，研究了这种简单的理论期望与来自人类种群的权威数据的符合程度。我们将祖先等位基因比例 $p_a(x)$ 与不同的 $x$ 频率制成表并且把它与预测 $p_a(x)=x$ 作了比较。

数据在预测直线的附近，但是斜率（0.83）明显的小于 1（图 32-4）。对于这种偏离的一个解释是一些祖先等位基因被确定下来的时候不正确。一个错误率 ε 会人为地将斜率降低 1～2ε。尽管如此，ε 值仅为 1.6%，这只能解释一小部分的偏离。最为可能的解释是人类历史中存在瓶颈，这使等位基因频率分布更平。理论计算指出近来的一个瓶颈会将斜率降低 $1-b$，$b$ 是瓶颈导致的近亲繁殖系数。这表明测量不同人类群体的斜率会揭示种姓特异性瓶颈。首先在一些 ENCODE 区域进行的等位基因频率分析（HapMap 协会 2005）表明欧洲和亚洲样本的斜率低于 1，非洲样本中接近 1，这与“走出非洲”人类迁移模型是一致的。

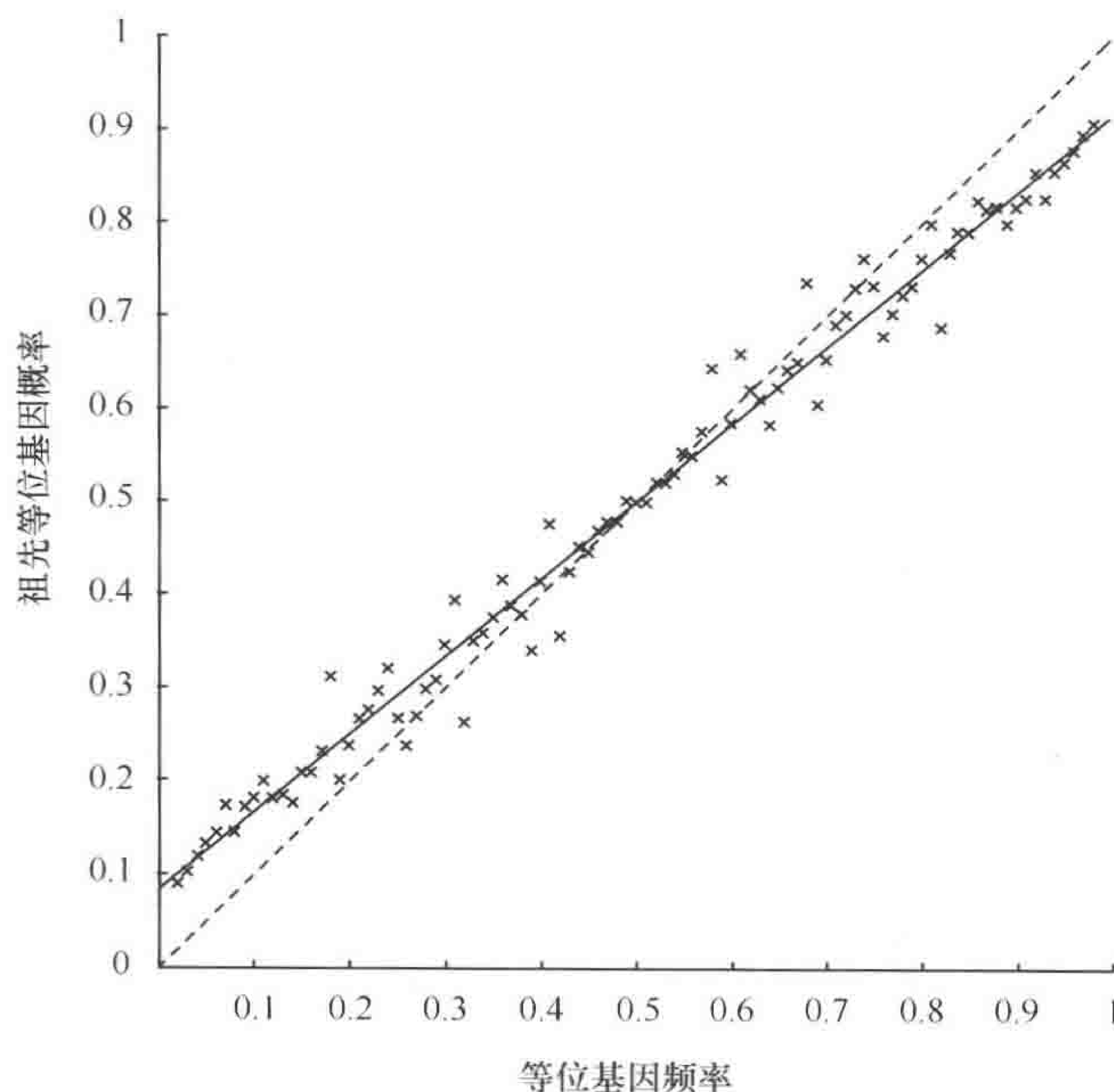



图 32-4　1%可观察频率中的可观察祖先等位基因。实线表示理论关系 $p_a(x)=x$。注意因为每个变种都产生一个衍生性的和一个祖先等位基因，数据必定要关于 0.5 对称（授权修改自黑猩猩测序和分析协会，2005）

## 自然选择的信号

测定黑猩猩基因组序列的一个主要动机就是阐明自然选择对人类进化历史的影响。过去的自然选择可以从那些与中性遗传漂变不同的序列分歧或变异模式中推测出来。通常将自然选择的影响分为两种主要形式：负的，或者说是纯化，选择导致一个有害的等位基因从种群中消失；正的，或者说是适应的，选择导致有利等位基因的快速固定。

现在已提出了无数种方法来检测负和正自然选择的信号。区别它们的一个重要方法是可以检测这些信号的时间框架。黑猩猩基因组序列在检测发生在5～7百万年的自然选择时间特别有用，这个时间分开了人类的LCA、黑猩猩种和解剖学上现代人的LCA（我们称这个时间跨度中的选择为古代选择）。结合人类遗传变异数据，黑猩猩序列可以提供一个有价值的基线来推断关于过去250 000年（近代）发生的自然选择。PAML（http://abacus.gene.ucl.ac.uk/software/paml.html）和MEGA（http://www.mega-software.net）是两个流行的软件包，可以执行此处描述的几种方法。

可以从功能分歧序列相对于中性进化的降低速率来辨别古代信号中的负选择。例如，从我们发生分歧以来平均蛋白质编码基因仅仅积累了两个氨基酸的替换，贯穿所有同源物的氨基酸改变（非同义）速率的平均比率和同义替代（通常表示为$K_A/K_S$或者$d_N/d_S$）是0.23，在没有自然选择时其期望值为1。在认为同义突变在选择上是中性的条件下，这意味着77%的氨基酸改变是有害的，足以在人类进化过程中被自然选择清除。这可能是一个较低的值，因为不断增加的证据表明在人类和其他哺乳动物中同义突变确定经历了一些负选择（Chamary et al. 2006）。使用共同的转座子元件的分歧速率，也可以在有功能的非编码序列（如5′和3′不翻译区及顺式调节元件）中检测到负选择作用。

可以通过蛋白质编码区域中的人类遗传变异模式来辨别近代信号中的负选择。检测频率大于15%的衍生等位基因HapMap SNP时，我们发现同义多态性氨基酸改变比率约为0.23，与固定下来的差异比率相似。这意味着负信号将大多数新的、有害的等位基因在达到较高频率之前就从现代人种群中除去了。尽管如此，具有1%～15%的衍生等位基因蛋白质编码基因的多态性确实有一个升高的$K_A/K_S$比率（从HapMap数据来估计为0.3～0.4）。这表明一系列的有害氨基酸改变可能临时迅速达到可检测水平的频率从而显著地贡献于人类遗传负担（Fay et al. 2001）。结合用软件，如Polyphen（Ramensky et al. 2002）和SIFT（Ng and Henikoff 2003）预测的具有功能相关替换的衍生等位基因频率数据有助于鉴定感兴趣的有害等位基因。

值得注意的是在人-黑猩猩分歧和人的多态性的$K_A/K_S$约为0.23，明显地高于其他得到很好研究的哺乳动物进行类似比较得到的结果。中性理论（Ohta 1998）预言这种增加表明了一种进化限制的一般松懈性，这是由于有效种群较小，因此相对其他一些哺乳动物来说遗传漂变对人类和黑猩猩的进化有更大的影响。

古代信号中的正选择比负选择更难以检测，但是意义更大。虽然有效中性等位基因具有表型效应，但一般认为正选择信号有助于精确地找到对人种进化历史最关键的遗传改变。

人类和黑猩猩间的低分歧率能有效地清除单纯通过统计平均数检测大多数基因或者其他功能元件中正选择的确切证据的可能性，但是一些方法已经被用于鉴定和排列用于后续分析的候选基因，并且阐明适应性进化的大尺度模型（Fay and Wu 2003）。这些方法都依赖相对于同义替换（$K_A/K_S$ 测试）的氨基酸过量改变的检测、在一种家系中相对于其他家系加快的替换率或者相对于人类群体中多态性数量过多的固定下来的替换的检测（the McDonald-Kreitman 测试）。用于检测非编码功能元件中正选择的类比方法的发展是更新的但却是更活跃的研究领域（Rockman et al. 2005）。

最严格的正选择信号是在蛋白质编码基因中氨基酸替换大大超过了同义替换，这不能用中性进化学说解释。$K_A/K_S$ 测试可以用于单家系，通过使用外群来推测家系特异性替换，也可以用于一对家系。后一种用法具有更强的统计学能力，如果一个基因在两个家系中都处于正选择之下。在我们刚开始的研究中，我们鉴定了人和黑猩猩中具有过多氨基酸替换的 585 个同源物，预计超过 50%都处于没有正选择的零假设之下。这一系列的同源物在涉及免疫系统的基因中含量十分丰富。到目前为止这是高等生物研究中发现的共同现象，反映了进化过程中对物种防御机制持续的选择压力。

标准 $K_A/K_S$ 测试的一个主要限制是它不能解释作用在一个基因上不同区域的异源的选择压力。对于蛋白质序列，将测试限制在特定的功能区域可能更有用。例如，对人类和黑猩猩 BRCA1 肿瘤抑制因子的正选择看起来集中在它的 RAD51 相互作用区域上（Huttley et al. 2005）。现在已经发展出了多种 $K_A/K_S$ 测试来检测局限于个别密码子的正选择（称作“异质性位点”和“分支位点”测试；Yang et al. 2000；Zhang et al. 2005），但是对这些方法在密切相关的物种的比较中的适用性还有争议。

替换速率具有家系特异性增加现象，这可以从人类、黑猩猩和一种或多种其他物种的比较中推断出来，它能提供更微妙的选择信号。限制是这个信号不能正确地区分正选择和进化限制松懈，并且它随着使用的外转群体不同而不同。例如，通过比较人和黑猩猩同源物中的 $K_A/K_S$ 值与小鼠和老鼠中对应的此比值，我们发现精子发生和雄性生殖系统中涉及的基因氨基酸替换表现出了高度的人特异性增加，潜在地反映了性选择对人种进化的强烈影响。相比之下，当使用鼠科作为外群时，人类中这些基因与黑猩猩相比并没有家系特异性增加现象，反映两个原始人家系具有更相似的选择压力。

值得注意的是，相对于猩猩，人转录因子基因确实显示出加速的氨基酸替换率。相对于现代人中的多态性氨基酸替代率，这些基因在人和黑猩猩间有明显过量的氨基酸替换（Bustamante et al. 2005）。这支持了基因规则的改变是我们的原始人祖先快速的解剖学进化中一个关键因素的假设（King and Wilson 1975）。

近代信号、正选择的意义也很大，因为它们是对查明解剖学意义上的现代人进化十分重要的基因，并且可以鉴定出对新病原体、食物和其他环境条件的遗传应答（Sabeti et al. 2006）。强烈的正选择产生了一个有特色的“选择性清扫”信号，因此一个稀少的等位基因迅速地固定了下来并且携带有这种现象发生频率很高的单体型（“搭便车效应”）。周围区域应该显示两种不同的信号：整体多样性显著减少和种群中产生过多的高频率衍生性等位基因，这是由于衍生性等位基因搭乘受选择性状的便车造成的（Przeworski 2002）。

黑猩猩基因组提供了精确估计两种信号所需的基线信息：当搜索低多样性的区域时可以将人-黑猩猩分歧作为突变率的中的区域变异的对照，也可以用黑猩猩基因组来将人的等位基因分为祖先性的或者衍生性的（上面已有讨论）。例如，在我们初始调查中，我们鉴定了 600 万个碱基区域，这个区域与人-黑猩猩分歧相比有着很低的人类多样性，这很特别，这里还有着过量高频率的衍生性等位基因，使得它们成为近代历史中强选择清扫作用的主要候选基因。在另一项研究中（Nielsen et al. 2005），根据 $K_A/K_S$ 测试，一系列基因被首先作为选择作用的候选基因鉴定了出来，研究中还发现它们具有过量的高频率的衍生性和氨基酸发生改变的基因，意味着适应性进化的近代事件。

# 结论

虽然由于道德限制和高昂的维护费用，黑猩猩作为一种传统意义上的模式生物还具有局限性，但是它的基因组序列仍是人类种群遗传学和人类家系进化研究中的重要资源。

# 致谢

我们感谢黑猩猩测序协会的全体成员，他们为第一次全面的人类和黑猩猩基因组的比较分析作出了重要贡献。我们还感谢 Manuel Garber，他制作了线性图。

# 参考文献

Altshuler D., Pollara V.J., Cowles C.R., Van Etten W.J., Baldwin J., Linton L., and Lander E.S. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407:** 513–516.

Bustamante C.D., Fledel-Alon A., Williamson S., Nielsen R., Hubisz M.T., Glanowski S., Tanenbaum D.M., White T.J., Sninsky J.J., Hernandez R.D., et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153–1157.

Chamary J.V., Parmley J.L., and Hurst L.D. 2006. Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7:** 98–108.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Ewing B. and Green P. 1998. Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Res.* **8:** 186–194.

Fay J.C. and Wu C.-I. 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet.* **4:** 213–235.

Fay J.C., Wyckoff G.J., and Wu C.-I. 2001. Positive and negative selection on the human genome. *Genetics* **158:** 1227–1234.

Glazko G.V. and Nei M. 2003. Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20:** 424–434.

Hellmann I., Prüfer K., Ji H., Zody M.C., Pääbo S., and Ptak S.E. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res.* **15:** 1222–1231.

Hellmann I., Zollner S., Enard W., Ebersberger I., Nickel B., and Pääbo S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13:** 831–837.

Hughes J.F., Skaletsky H., Pyntikova T., Minx P.J., Graves T., Rozen S., Wilson R.K., and Page D.C. 2005. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437:** 100–103.

Huttley G.A., Easteal S., Southey M.C., Tesoriero A., Giles G.G., McCredie M.R., Hopper J.L., and Venter D.J. 2000. Adaptive evolution of the tumor suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. *Nat. Genet.* **26:** 131–132.

International HapMap Consortium. 2005. A haplotypes map of the human genome. *Nature* **437:** 1299–1320.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Jaffe D.B., Butler J., Gnerre S., Mauceli E., Lindblad-Toh K., Mesirov J.P., Zody M.C. and Lander E.S. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13:** 91–96.

King M.C. and Wilson A.C. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116.

Kuroki Y., Toyoda A., Noguchi H., Taylor T.D., Itoh T., Kim D.S., Kim D.W., Choi S.H., Kim I.C., Choi H.H., et al. 2006. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat. Genet.* **38:** 158–167.

Newman T.L., Tuzun E., Morrison V.A., Hayden K.E., Ventura M., McGrath S.D., Rocchi M., and Eichler E.E. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15:** 1344–1356.

Ng P.C. and Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31:** 3812–3814.

Nielsen R., Bustamante C., Clark A.G., Glanowski S., Sackton T.B.,

Hubisz M.J., Fledel-Alon A., Tanenbaum D.M., Civello D., White T.J., et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3:** e170.

Ohta T. 1998. Evolution by nearly-neutral mutations. *Genetica* **102–103:** 83–90.

Patterson N.P., Richter D.J., Gnerre S., Lander E.S., and Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441:** 1103–1108.

Przeworski M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* **160:** 1179–1189.

Ramensky V., Bork P., and Sunyaev S. 2002. Human non-synonymous SNPs: Server and survey. *Nucleic Acids Res.* **30:** 3894–3900.

Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Dichter D.J., Lavery T., Kouyoumjian R., Farhadian S.F., Ward R., and Lander E.S. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199–204.

Rockman M.V., Hahn M.W., Soranzo N., Zimprich F., Goldstein D.B. and Wray G.A. 2005. Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biol.* **3:** e387.

Rosenberg H.F. and Feldmann M.W. 2002. The relationship between coalescence times and population divergence times. In *Modern developments in theoretical population genetics: The legacy of Gustave Malécot* (ed. M. Slatkin and M. Veuille), pp. 130–164. Oxford University Press, Oxford, United Kingdom.

Sabeti P.C., Schaffner S.F., Fry B., Lohmueller J., Varilly P., Shamovsky O., Palma A., Mikkelsen T.S., Altshuler D., and Lander E.S. 2006. Positive natural selection in the human lineage. *Science* **312:** 1614–1620.

Schwartz S., Kent W. J., Smit A., Zhang Z., Baertsch R., Hardison R.C., Haussler D., and Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.

Warren R.L., Varabei D., Platt D., Huang X., Messina D., Yang S.P., Kronstad J.W., Krzywinski M., Warren W.C., Wallis J.W., et al. 2006. Physical map-assisted whole-genome sequence assemblies. *Genome Res.* **16:** 768–775.

Watanabe H., Fujiyama A., Hattori M., Taylor T.D., Toyoda A., Kuroki Y., Noguchi H., BenKahla A., Lehrach H., Sudbrak R., et al. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429:** 382–388.

Watterson G.A. and Guess H.A. 1977. Is the most frequent allele the oldest? *Theor. Popul. Biol.* **11:** 141–160.

Yang Z., Nielsen R., Goldman N., and Pedersen A.M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.

Yohn C.T., Jiang Z., McGrath S.D., Hayden K.E., Khaitovich P., Johnson M.E., Eichler M.Y., McPherson J.D., Zhao S., Pääbo S., and Eichler E.E. 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLoS Biol.* **3:** e110.

Yunis J.J. and Prakash O. 1982. The origin of man: A chromosomal pictorial legacy. *Science* **215:** 1525–1530.

Zhang J., Nielsen R., and Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22:** 2472–2479.

Zuckerkandl E. and Pauling L. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8:** 357–366.

# 33　系谱标记：mtDNA 和 Y 染色体

Mark Stoneking[1] and Manfred Kayser[2]

[1]*Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany;* [2]*Department of Forensic Molecular Biology, Erasmus University Medical Centre, 3000 CA Rotterdam, The Netherlands*



## 简介

由于线粒体 DNA（mtDNA）和 Y 染色体上非重组区都是单倍体，遗传自一个亲本，不发生任何重组，它们提供了并且还在提供人类群体或个体系谱历史的重要信息。此外，由于每个细胞中都有多个拷贝，mtDNA 适用于分析远古 DNA 基因组。本章将讨论由 mtDNA 和 NRY 变异分析引出的一般性课题，还将拿出我们自己在人类群体 mtDNA 和 NRY 变异的比较分析方面所做的一些工作。最后，简要总结 mtDNA 和 NRY 标记在法律证明和系谱研究中的运用。

## mtDNA 变异

使 mtDNA 适用于群体和进化遗传学研究的特性包括多拷贝、快速的进化速率、单倍体、单亲遗传、遗传过程中没有重组等，对此已经有了详细的总结（Pakendorf and Stoneking 2005）。在 PCR 发明之前，曾经用 Southern blot 方法对全基因组 DNA 进行低分辨限制性位点作图（Johnson et al.　1983），或使用纯化的末端标记 mtDNA 进行高分辨作图（Cann et al.　1987）进行人类 mtDNA 变异的研究。1981 年确定的人类线粒体基因组的全序列（Anderson et al.　1981）使这两项研究能够根据所观察到的 RFLP 推断一些突变的存在。PCR 技术出现后，一些研究者使用 PCR 片段的高分辨作图来进行人类 mtDNA 变异研究（Torroni et al.　1992）。

序列研究显示mtDNA上能够提供信息的多数变异来源于非编码的调控区，其中绝大多数出自一个称为HV1（hypervariable segment）的区域。目前许多人类mtDNA变异的研究集中于对HV1序列的分析上，已经公布了上万个HV1序列。除了像Genbank这样的DNA序列数据库外，Hvrbase＋＋（www.hvrbase.org）也对HV1序列进行了有益的编辑。最近，随着序列测定的时间和费用的降低，研究工作已转而对mtDNA基因组全序列进行测定（Ingman et al.　2000），得到了mtDNA所能提供的最大信息。一些研究已经提供了特殊mtDNA世系的重要的、令人感兴趣的历史资料（Thangaraj et al.　2005；Trejaut et al.　2005），但在目前，测试时间和费用的限制使得群体遗传研究所需要的大规模mtDNA基因组全序列测序仍然受到限制。

虽然HV1序列作为标准的分析人类mtDNA变异的方法被数十家实验室所采用，但HV1也有自身缺陷。HV1包含一个起始于核苷位置（np）16 024（控制区域的起点）的大约380 bp的区域（Anderson et al.　1981），这是一个便于扩增和做序列分析的长度。然而在np16 189的一个T→C突变（或缺失）导致被称为"C-延展"（C-stretch）的9～10个连续的胞嘧啶的出现，在PCR扩增时由于这一序列过长而可能造成滑动（Bendall and Sykes　1995）。因此有必要或者每段测序两次，或者使用内部的引物来避开C-延展区域，以使测序正确。

在过去几年中，对HV1序列的错误已经有许多讨论（Bandelt et al.　2002；Forster　2003；Salas et al.　2005）。已经有一些分散的实验室专门检查已经公开的研究和数据库的错误。目前是通过对当前所观察序列的变异类型的统计分析方法来发现错误（Bandelt et al.　2002），但由于新的序列不能肯定过去的类型，这种分析也可能有潜在的问题（Barbujani et al.　2004），况且这样的统计分析不能涵盖所有潜在的错误，因此不能够代替仔细的实验室工作。在这样一个情形下，我们懊丧地注意到在公开发表的研究中，只从HV1 PCR产物中的一条链得到序列信息的倾向在增加。我们的观点是常规地测定两条链是绝对需要的，因为以我们的经验看，这是检出错误的最好的方法（特别是在样品混淆时，这也是最常见的错误）。尽管并不十分简洁，为发现序列错误，将新的HV1序列与已经有的序列进行比较找出差异类型，是一项重要的且有用的方法，应该应用于每项研究中。

早期关于mtDNA RFLP变异的一个研究成果是可以根据共同的确认位点将个体mtDNA类型归入不同的单倍组，这些单倍组显示出一定程度的地域特异性。主要的单倍组按照字母表的顺序排列，但由于缺乏命名新的单倍组的统一的命名体制，单倍组树的构造相当随意。目前单倍组树的版本可以在Mitomap Web Site上找到（www.mitomap.org），主要单倍组的地理分布图也可以在这个网站找到。一些单倍组可以从HV1序列中推导出来，其可信度很高，但有一些不行。对于后者来说，确认的位点必须进行分型以便归入单倍组。

分析mtDNA序列和（或）单倍组资料时［或者NRY单倍型和（或）单倍组资料时，后面将讨论到］将用到以世系为基础的或以群体为基础的两种基本方法（Pakendorf and Stoneking　2005）。世系为基础的方法按种系发生地理学的原则分析每一个清楚的mtDNA世系（如单倍组），通常按每一个世系的起源时间和地理分布来限定。这

种研究可以得到对 mtDNA 世系历史的详细了解，但是不能直接揭示携带这些世系的群体的历史。不幸的是，这种以世系为基础的分析有局限性的观点在开始时没有被广泛接受，当时有一种天真地将单倍组的年龄与群体的年龄等同起来的倾向，并且假定群体中每个单倍组反映了各自的迁徙过程。幸好，现在已经广泛地接受对于 mtDNA 和 NRY 的数据需要用群体遗传学的方法去推断群体的历史的观点；这种方法使用群体，而不是单个世系，作为分析的基本单位。作为运用这种方法开展工作的一个例子，我们在此介绍如何根据 mtDNA 和 NRY 的数据，洞察玻利尼西亚人的殖民过程。

有两个论题在质疑将 mtDNA 分析应用于人类群体的变异和进化的合法性：①重组；②选择。关于重组，基于相当有限的系谱研究（Giles et al. 1980)，多年流行的观点是 mtDNA 属于严格的母性遗传，不会发生重组。这种观点受到了一些研究的挑战。这些研究根据一些 mtDNA 序列中观察到的不寻常的变异类型（Hagelberg et al. 1999）和变异类型的统计分析（Awadalla et al. 1999）提出了重组的证据。然而，事后证明这些不寻常的变异类型是人造的（Hagelberg et al. 2000)，并且那些统计分析也是有缺陷的（Kumar et al. 2000)。尽管在一个 mtDNA 疾病家庭中有一个个体表现出人类 mtDNA 的父性遗传（Schwartz and Vissing 2002)，并且在这一个体的体细胞里发现了重组（Kraytsberg et al. 2004)，但成千的母亲-后代的比较分析都未能提供发生过重组的证明（Pakendorf and Stoneking 2005)。一些变异类型原被认为提示了重组，但较一致的观点认为在高变位点发生的平行或回复突变是观察到的变异类型更好的解释（Piganeau and Eyre-Walker 2004)。

类似地，有人曾经提出对可能与气候适应有关的特定 mtDNA 突变的自然选择能够说明在特定群体中 mtDNA 的分布类型（Excoffier 1990；Mishmar et al. 2003；Ruiz-Pesini et al. 2004)，这造成对使用 mtDNA 推断群体历史的怀疑。然而，尽管作为一种纯化选择的非中性进化（Nachman et al. 1996）和群体的生长（Rogers and Harpending 1992）影响到人类 mtDNA 变异的类型，适应性进化的主张并没有被证实（Kivisild et al. 2006；Sun et al. 2006)。在任何情况下，任何基于 mtDNA 作出的群体历史的推论应该通过对新的位点的分析来证实，由于选择或偶然事件，仅根据单一位点作出的分析总是有风险的，有可能产生对群体历史的错误认识。

## NRY 变异

人类的 Y 染色体大约有 23 000 000 bp，作为我们基因组中最大的单倍性染色体，可以逃脱染色体内的重组，它本应能提供男性系谱研究的丰富的特异性标记。然而很长时间它被认为缺乏有用的 DNA 多态性。直到 1985 年，第一个 NRY 标记：12f2 或 DYS11 才开始使用（Casanova et al. 1985)。经过了 20 多年对有用标记的广泛研究，现在已可以利用一组详尽了解的 NRY 标记。针对世界范围内提供的 DNA 样品已分析了总共 245 个 NRY 单核苷酸多态性（Y-SNP)，由此建立广泛的系谱树，连同首先被 Y 染色体协会（YCC 2002）采纳的，通常使用的基于 Y-SNP 的单倍组的命名体制，越来越多的标记被发现并加入其中（Jobling and Tyler-Smith 2003)。大多数标记是通

过将变性高效液相色谱（dHPLC）应用于世界范围内的样本的短 Y 染色体扩增片段，结合通常从非洲来的对照 DNA 而被发现的（Underhill et al. 2000）。由于在标记排查过程中使用的个体数目有限且经过挑选，这个步骤造成了一种严重的探索偏差。当 Y-SNP用于地理区域多样性估计时需要特别留意这些偏差，在最初的排查中即样品不足。

除了 Y-SNP 外，在男性系谱研究中，成百的 NRY 短串连重复多态（Y-STR）或微卫星已经被发现并被用来补充 Y-SNP 的结果。1992 年，利用传统的分子技术，即重复探针与 Y 染色体克隆杂交的方法，发现了第一个 Y-STR：27H39LR 或 DYS19（Roewer et al. 1992）。另外几个 Y-STR 也这样被确认和鉴定（Kayser et al. 1997）。然而更多的 Y-SNP 是作为国际人类基因组计划的一部分，通过对几乎完全的人类 Y 染色体 DNA 序列测定而于最近被发现的。利用这一资源，使用能够认识单重复序列结构的计算机运算，随后通过分子技术确认标记，166 个以前未知的、有用的 Y-STR 被鉴定出来（Kayser et al. 2004）。

与常染色体的 STR 相比，Y-STR 的复杂性之一在于 Y 染色体的高度重复性和回文结构（Skaletsky et al. 2003），一些 Y-STR 位点是多拷贝的，这意味着在使用一对引物进行 PCR 扩增后会出现两个（或多个）不同的男性特异性的扩增产物。在通常的重复位点的情况下，这代表了两个位点，如 DYS385a/b，这种情况并不少见（Kayser et al. 2004）。然而有时重复位点只产生单一产物，在这种情况下通常假定两个位点的等位片段有相同的长度，虽然原则上有可能其中一个位点缺失。在将各等位片段指派到个体的一对重复位点时，习惯上假定较大的等位片段属于一个位点，而较小的一个属于另一个位点，但这并不能应用于所有情况，就像一种分开 DYS385a/b 位点的方法发展起来后所显示的那样（Kittler et al. 2003）。另一个有时出现的复杂因素是可能发生新的重复，导致重复位点存在或不存在的多态现象。例如，DYS19 的重复在蒙古人和蒙古起源的群体中很常见，如 Kalmyk 人（一个位于美国西部的属于蒙古人种的种族）（Zerjal et al. 2003）。

按照惯例，单倍组被用来指示一个个体的 Y-SNP 的状态，而单倍型是指 Y-STR（de Knijff 2000）。应注意这个术语与应用于常染色体变异的术语不同，在常染色体中单倍型可以由 STR 和 SNP 的任意组合组成。产生这种区别的原因是由于不存在重组，Y-SNP 代表独特的事件，它标记着独特的 Y 染色体世系。沿着常染色体的重组意味着单个常染色体 SNP 不能够标记独特的世系。虽然原则上 STR 和 Y-SNP 在这方面是一样的，但实际上在 Y-STR 位点的高频率的平行和回复突变意味着具有同样 Y-STR 单倍型的个体可能属于不同的 Y-SNP 单倍组。正常情况下，Y-SNP 被用来鉴定更早期的系谱事件，如数千年前；而 Y-STR 应用于群体中近期发生的事件或个体遗传历史。这是因为 Y-STR 的突变率比 Y-SNP 高 100 000 倍（Kayser et al. 2000b；Thomson et al. 2000）。所以用远古的事件常常被近期发生的 Y-STR 突变所蒙蔽，使得当只有 Y-STR 被分析时，它们的鉴定工作变得困难（Kayser et al. 2001）。因此今天以 NRY 为基础的谱系研究通常运用这两种 Y 染色体多态性，以便取长补短，充分地获得 Y 染色体提供的信息：Y-SNP 用于根据父系所作的个体或群体的宽泛分类（也就是 NRY 单倍

组）；Y-STR 用来确定谱系起源的地点和时间。这样的结合可以深入了解男性系谱，并提供关于人类（男性）群体的历史（Zerjal et al. 1997，2003；Kayser et al. 2000a；Seielstad et al. 2003）。在一些实例中，单倍组和单倍型可得出相似的结论（Rosser et al. 2000；Semino et al. 2000；Roewer et al. 2005），并在极少的情况下提供基本一致的信息（Kayser et al. 2005）。

由于 NRY-DNA 的完全连锁，对人类 Y 染色体的自然选择将明显地影响整个染色体的遗传变异。这一论题已在别处讨论过（Jobling and Tyler-Smith 2003），所得到的结论是：基于现有的知识，没有证据显示其作用在人类 Y 染色体的差异选择。然而已知一些影响生殖成功的文化因素也影响到 Y 染色体的多样性。其中一种文化因素是婚后到男方家居住，下面将要讨论到这一点。其他影响 Y 染色体多样性的文化因素还包括一夫多妻（一个男性与许多女性养育子女），这是许多在基督教之前或基督教之外的人类社会的组成部分；还有涉及选择性地屠杀男性以便获得妇女的战争，这在某些地区曾经实施过，如新几内亚。婚后到男方家居住、一夫多妻和屠杀男性的战争都导致 Y 染色体多样性的减少和 Y 染色体的非均匀分布。这些因素可以解释在西巴布亚新几内亚人类群体中在传统条件下所观察到的 Y 染色体，而不是 mtDNA 的多样性减少，这个群体直到最近还存在着，其中一部分延续到今天（Kayser et al. 2003）。极端的偏于男性的迁徙过程，如作为战争性质的入侵所导致的一种结果，也会反映在 Y 染色体的多样性上。一个例子是，一个在整个中亚有相对较高的频率（大约 8%）并且有特殊分布的紧密相关的 Y 世系簇，反映了成吉思汗和他的后代的遗传足迹（Zerjal et al. 2003）。

## 人类进化

在当代人类群体中，mtDNA 变异最初被解释为支持人类 mtDNA 的近期非洲起源（Cann et al. 1987；Vigilant et al. 1991），并且后来的工作进一步强化了这一解释（Pakendorf and Stoneking 2005）。下述事实都证明了这一点：①非洲群体具有更多的 mtDNA 变异（通常情况下，具有最多数量的突变的群体最可能是祖先群体）；②mtDNA类型的系统发生树总是包括两个最初分支，其中一个肯定由非洲的 mtDNA 类型构成，另一个 mtDNA 类型来源于非洲和世界其他地区（如果共同的祖先在非洲的话，这一模式就非常容易解释）；③大约 150 000 年前的人类 mtDNA 祖先的估计年代和对大约 50 000 年前 mtDNA 扩张到非洲之外的变异模式的进一步阐明（Ingman et al. 2000）。这些基于对现代群体研究的结论进一步为尼安德特人（Krings et al. 1997）和早期现代人（Serre et al. 2004）的 mtDNA 序列分析结果所支持：尼安德特人 mtDNA 序列被排除在现代人变异的范围之外，尼安德特人的 mtDNA 没有与古代人或现代人的任何关联。

许多年来，涉及现代人起源的遗传证明的讨论集中在 mtDNA 的资料上。当 NRY 上的资料可供利用并且提示了人类 NRY 变异近期起源于非洲时令人鼓舞（Underhill et al. 2000）。然而 NRY 祖先的年龄较 mtDNA 祖先的年龄相对年轻，为 60 000～

100 000年（Macpherson et al. 2004）。总体上来讲，在产生子女和后代方面，人类的男性群体有比女性群体小的倾向，这一显著的偏差实际是可以预期的。

人类 mtDNA 和 NRY 变异近期的非洲起源并不意味着我们全部基因组的近期非洲起源，这一点需牢记在心。一些研究确实指出我们基因组中的一小部分可能来源于与原始人类的混杂（Wall and Hammer 2006）。尼安德特人（及其他原始人类）的基因组 DNA 的测序应该能提供这一问题的确切答案（Green et al. 2006；Noonan et al. 2006）。

## 不同人类群体 mtDNA 和 NRY 变异的比较

早期对不同人类群体 mtDNA 和 NRY 变异的比较分析结果显示：NRY 的群体之间的遗传学距离平均地要大于 mtDNA 的遗传学距离（Seielstad et al. 1998）。群体之间遗传的区别反映了遗传漂移（从一个群体到另一个群体基因频率的随机变化）和群体间迁徙或基因流动的平衡，前者造成群体之间在遗传上的相异，后者引起群体之间遗传相似性的维持。因此对于 NRY 较大的遗传距离可被解释为上述的人类群体中男性相对于女性较低的迁徙率。全球范围内男性较女性低的迁徙率的程度近期受到挑战（Wilder et al. 2004），但在局部水平上，NRY 的群体间的遗传距离总是大于 mtDNA 的遗传距离（Kayser et al. 2003；Nasidze et al. 2004；Pakendorf et al. 2006）。

由于我们习惯地认为迁徙主要是男性的入侵和征服，因此男性迁徙率小于女性的想法看起来违反直觉。然而 mtDNA 和 NRY 的资料可能从另一方面反映了居住类型效应（Seilestad et al. 1998）：多数人类群体是婚后居住在男方家里的，一旦结婚，女性将转移到男方的居住地。婚后居住在男方处所这一事实说明了男方在不同的人群之间移动性少于女方，由此可以预见所观察到的群体之间较大的 NRY 遗传距离。对这一假说的检验方法应该是检查婚后居住在女方处所群体的 mtDNA 和 NRY 的变异类型，预期结果应该是这样的群体显示相反的变异类型，即 mtDNA 的遗传距离较 NRY 为大。一个关于泰国山区部落中的婚后居住在男方处所群体和婚后居住在女方处所群体的研究结果确实如此，令人惊讶（Oota et al. 2001），这一预见得到了相当满意的证实（图33-1）。

这个例子表明当研究者正在考察不同的社会或文化形态下男性和女性行为时，对 mtDNA 和 NRY 变异的比较分析是特别可信的。另一个例子是关于印度以前的等级制度，这一制度造成了男女之间的婚姻选择的各种社会限制，导致了在等级制度之内和之间的对 mtDNA 和 NRY 变异结构的预期判断。这一预期又一次被精确证实（Bamshad et al. 1998）。如果没有 mtDNA 和 NRY 的分析，我们不可能达到对这一制度及其他社会和文化现象对人类群体遗传结构的影响的如此洞察力，并且就这一理由来讲，它们属于各种系谱标记中对人类群体遗传学最具显著的贡献者之列。

## 研究事例：玻利尼西亚人的起源

在对特定群体的 mtDNA 和 NRY 变异进行分析时，多数情况下可假定人类群体的母源和父源历史相同。然而有时他们却是不同的，也许最富于戏剧性的例子是玻利尼西

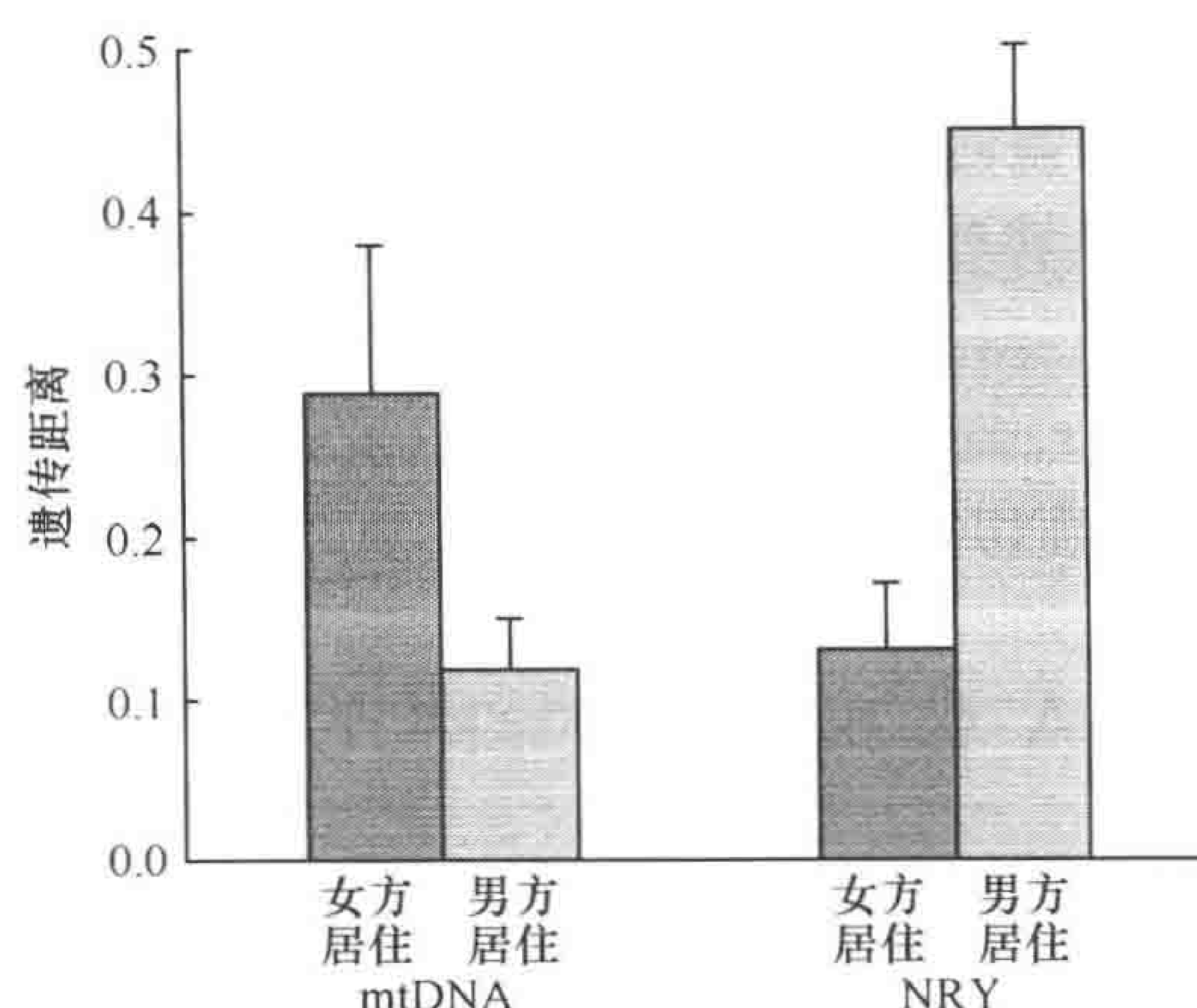



图 33-1　泰国山区部落中的婚后居住在男方处所群体和婚后居住在女方处所群体的遗传距离。在婚后居住在男方处所群体显示 NRY 通常比 mtDNA 有更大的遗传区别类型；而在婚后居住在女方处所群体这一情形正相反。(Oota et al.　2001)

亚人的起源。

玻利尼西亚人是太平洋许多岛屿（玻利尼西亚）的居住者，分布在一个广大的三角地带，西接图瓦卢，北临夏威夷，东面是复活节岛，南面是新西兰。由于这些岛屿相距甚远（距离大陆有数千千米之遥），至少从 16 世纪他们被第一批欧洲海员发现后，玻利尼西亚人的起源和迁徙历史不仅让科学家们十分感兴趣，同样也使一般人感兴趣。依据考古学的资料，3000～800 年前玻利尼西亚开始被人类所占据，西玻利尼西亚较东玻利尼西亚更早出现人类，而新西兰人是最后到达的（Kirch　2000）。考虑到澳大利亚和巴布亚新几内亚在 50 000～40 000 年前就被人类占据（Groube et al.　1986；Roberts et al.　1990），玻利尼西亚这么晚才出现人类是令人惊讶的。有两个关于玻利尼西亚人起源的极端的假说，其中“快车”模型（“Express train” model）假定大约 6000 年前他们从东亚起源，然后迅速向东进入太平洋地区，没有明显的停滞。这种停滞可导致和附近群体的混杂，如和新几内亚土著居民混杂（Diamond　1988）。这一模型预言玻利尼西亚人的特征主要来自（如果不是全部）亚洲。而另一个“杂乱河堤”模型（“entangled bank” model）则与此相反，它假定在新几内亚和包括玻利尼西亚在内的周围地区有一个长时间的相互作用，其过程从人类最开始定居新几内亚即已开始（Terrell　1988）。因此这一模型预测玻利尼西亚人的特征主要来自（如果不是全部）新几内亚人或美拉尼西亚人（这里指的是一个地理区域，包括新几内亚大陆和周围岛屿）。其他居于中间的模型，如“走出台湾”模型（Bellwood　2004）或“三重-I”模型[“Triple-I” model 指的是 intrusion（入侵）、innovation（改造）、integration（整合）三个词的首字母，Green　1991]也假定玻利尼西亚人祖先起源于东亚，并允许在玻利尼西亚人祖先和美拉尼西亚人之间有一定量的接触和混合（未详细说明）。按照这些居间模型，玻利尼西亚人应该具有亚洲人和美拉尼西亚人两者的特征。

为澄清关于玻利尼西亚人起源的不同的假说，需要评估亚洲人和美拉尼西亚人对玻利尼西亚人相对的遗传贡献。有两个 mtDNA 和 NRY 单倍组的特征可被用来推断他们

的地理起源：①地理分布，②相关的突变多样性的量。一般来讲，可以期待一个单倍组起源于这样一个群体，该群体呈现最大量的相关突变多样性，由于缺乏重组，突变的多样性主要是时间的函数。可以容易地将玻利尼西亚人的 mtDNA 和 NRY 单倍组分为两种类型：一类在整个东亚和东南亚、新几内亚的沿海地区（不包括高地）和美拉尼西亚岛屿均广泛存在；另一类在美拉尼西亚岛屿、新几内亚沿海和高原地区及东印度尼西亚被发现，而在东亚或东南亚其他地方不存在（Kayser et al. 2006）。第一种类型的 mtDNA 和 NRY 单倍组是亚洲起源的，而第二种则是美拉尼西亚起源的，这些推论由与这些单倍组相关的突变多样性所证实。亚洲组具有与第一类 mtDNA 和 NRY 单倍组相关的最高的突变多样性，而美拉尼西亚组具有与第二类单倍组相关的最高的突变多样性（Kayser et al. 2006）。图 33-2 显示跨越亚洲/太平洋地区的这种频率分布；图 33-3 显示亚洲单倍组 B4a（mtDNA）和 O-M122（NRY），以及美拉尼西亚人单倍组 Q1（mtDNA）和 C-M208（NRY）地区多样性的差异。

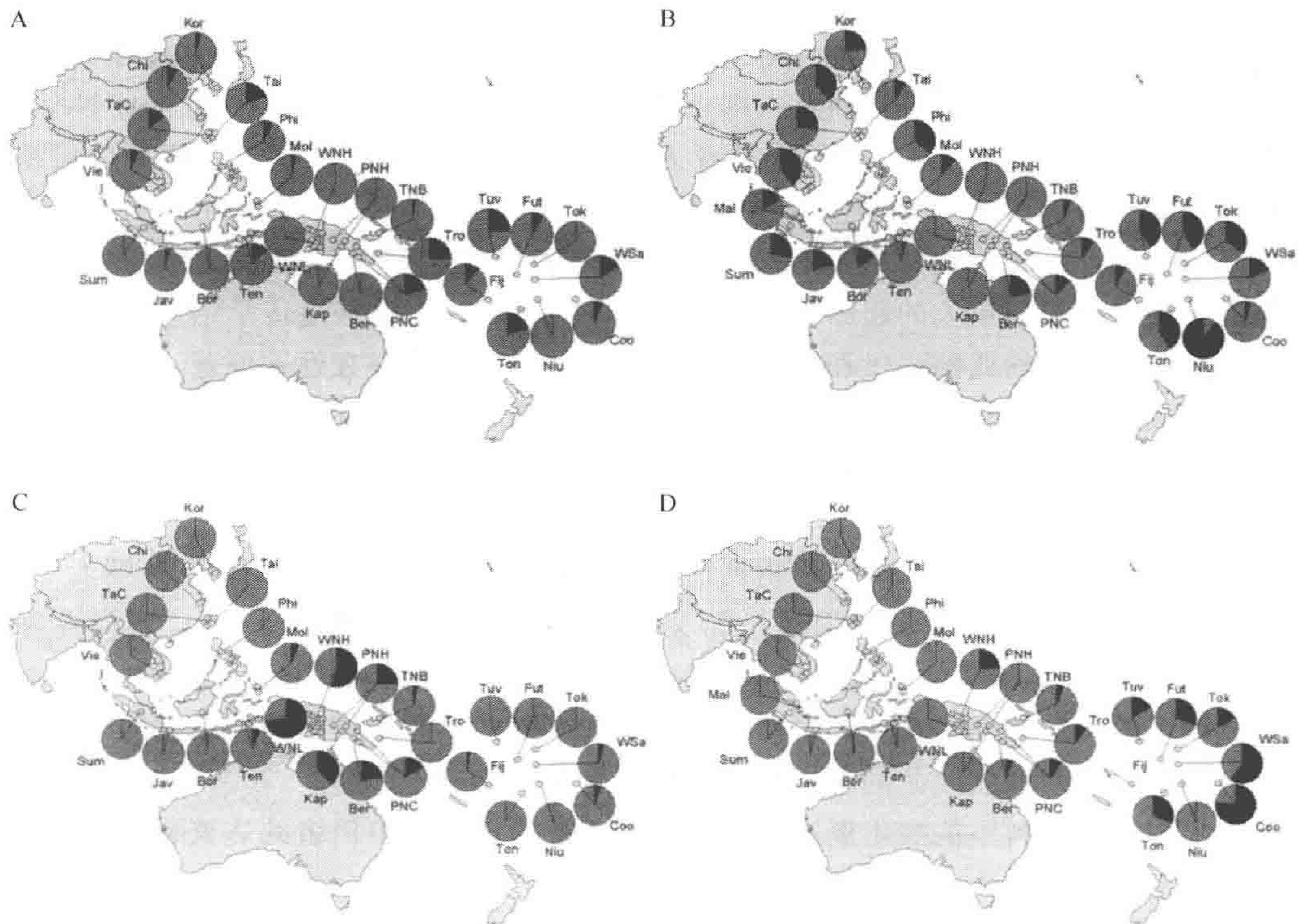



图 33-2 亚洲人和美拉尼西亚人单倍组和它们在整个亚洲/太平洋地区的频率分布的例子。A. 表示亚洲人 mtDNA 单倍组 B4a；B. 表示亚洲人 NRY 单倍组 O-M122；C. 表示美拉尼西亚人mtDNA 单倍组 Q1；D. 表示美拉尼西亚人 NRY 单倍组 C-M208. 饼图中黑色的部分代表所分析的群体样品中各自的单倍组频率。群体的缩写：库克（Coo）、纽埃岛（Niu）、托克劳群岛（Tok）、西萨摩亚（WSa）、汤加（Ton）、东富图纳（Fut）、土瓦鲁（Tuv）、斐济（Fij）、新不列颠（TNB）、巴布亚新几内亚所属特罗布里恩群岛（Tro）、巴布亚新几内亚所属贝雷纳（Ber）、巴布亚新几内亚海岸（PNC）、巴布亚新几内亚高地（PNH）、巴布亚新几内亚所属卡普纳（Kap）、WNG 高地（WNH）、WNG 低地（WNL）、摩鹿加（Mol）、田格拉斯岛（Ten）、菲律宾（Phi）、南婆罗洲（Bor）、爪哇（Jav）、苏门答腊（Sum）、马来西亚（Mal）、朝鲜（Kor）、中国台湾原住民（Tai）、中国台湾外迁人（TaC）、越南（Vie）(Kayser et al. 2006)

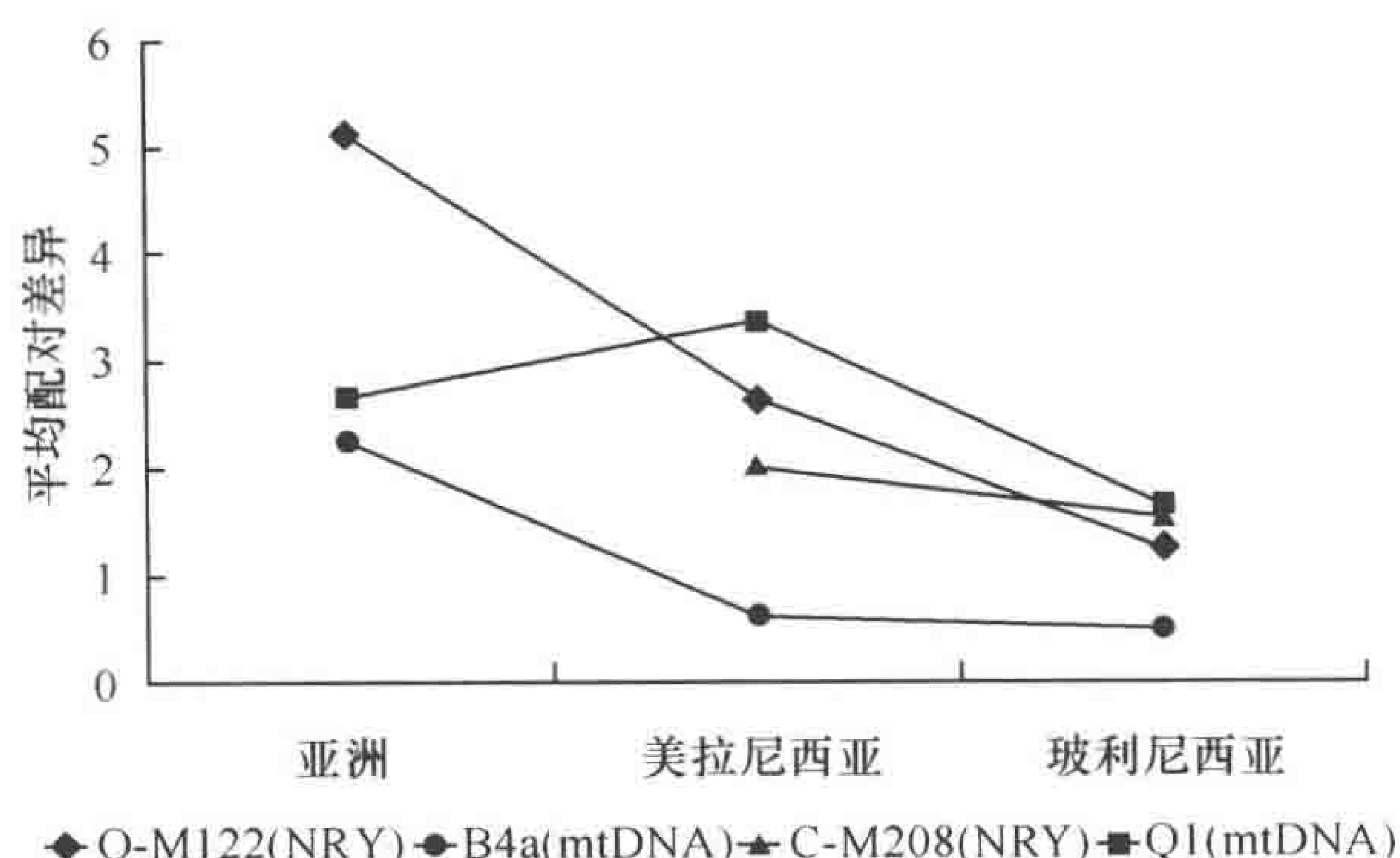



图 33-3　NRY 单倍组 C-M208 和 O-M122 的 Y-STR 相关多样性和与 mtDNA 单倍组 B4a 和 Q1 相关的 HVl 序列单倍体多样性的地区比较。多样性以配对比较差异的平均数来表示（Kayser et al.　2006）

引人注目的是，估计的亚洲人对 mtDNA 单倍组的贡献比估计的亚洲人对 NRY 单倍组的贡献大得多，而美拉尼西亚人则正相反。结合从亚洲人起源的 5 个 mtDNA 和 7 个 NRY 单倍组的证据，以及美拉尼西亚人起源的 4 个 mtDNA 和 7 个 NRY 单倍组的证据，我们可以确定 94％的玻利尼西亚人的 mtDNA 和仅仅 28％的玻利尼西亚人 Y 染色体来自亚洲，66％的玻利尼西亚人的 Y 染色体和仅仅 6％的玻利尼西亚人的 mtDNA 可被追溯到美拉尼西亚人的起源（Kayser et al.　2006）。对这些遗传学结果的一个解释是，到达新几内亚的亚洲移民和新几内亚原居民的遗传混合存在性别偏差，较多的亚洲妇女（和较少的男人）和较多的美拉尼西亚男人（和较少的妇女）参与混合。假定的远古时期的玻利尼西亚母系世系社会和婚后居住在女方住所将支持这种情形（Hage　1998）。

亚洲人和美拉尼西亚人的单倍组到达玻利尼西亚是一次还是几次迁徙浪潮的结果？这一问题可以原则上地被表述为通过相关的 mtDNA HVl 序列和 Y-STR 多样性确认 mtDNA 和 NRY 单倍组到达玻利尼西亚的时间。由于当今可利用的定时方法不可避免地造成较大的可信区间，在玻利尼西亚人中所观察到的不同单倍组的到达时间与亚洲人起源的和美拉尼西亚人起源的单倍组之间相互重叠（Kayser et al.　2006）。考虑到这种情形，根据现有资料不能找出亚洲人单倍组和美拉尼西亚人单倍组到达玻利尼西亚的时间上的差别。

mtDNA 和 NRY 标记可以有两种资料分析方法：基于群体的方法和基于世系的方法。在基于群体的分析中，群体是分析的单位，关注点是一个群体和另一个群体的关系。当进行基于群体的分析时，如用 NRY 单倍组频率或 mtDNA 序列所建立的 Fst 值（来源于亚群体与整个群体比较结果的固定指数）的多维标度，玻利尼西亚各群体聚合在一起，他们在 NRY 和 mtDNA 标记上与美拉尼西亚人、东亚和东南亚群体稍有分离（图 33-4）。群体根据地理区域的聚合也见于东亚/东南亚群体和美拉尼西亚各群体，岛屿美拉尼西亚人/沿海新几内亚人与新几内亚高地群体/西新几内亚人相互分离。斐济人是美拉尼西亚人中与玻利尼西亚人的聚合最近的。在 NRY 地区图上，由于美拉尼西亚

NRY 单倍组在东印度尼西亚的高频率，东印度尼西亚群体（摩鹿加和田格拉斯岛）与岛屿美拉尼西亚人和沿海新几内亚人组合在一起，在这些样品中这一频率高于美拉尼西亚人 mtDNA 单倍组。根据 mtDNA 得到的遗传距离和多维标度（MDS）地区图之间比根据 NRY 得到的有更好的适合程度，这可以用应力值来表示的（mtDNA：0.051 对比 NRY：0.157）。应力值提供了地区图与不同群体之间观察距离吻合度的测量指标，可由产生 MDS 地区图的程序计算出来。

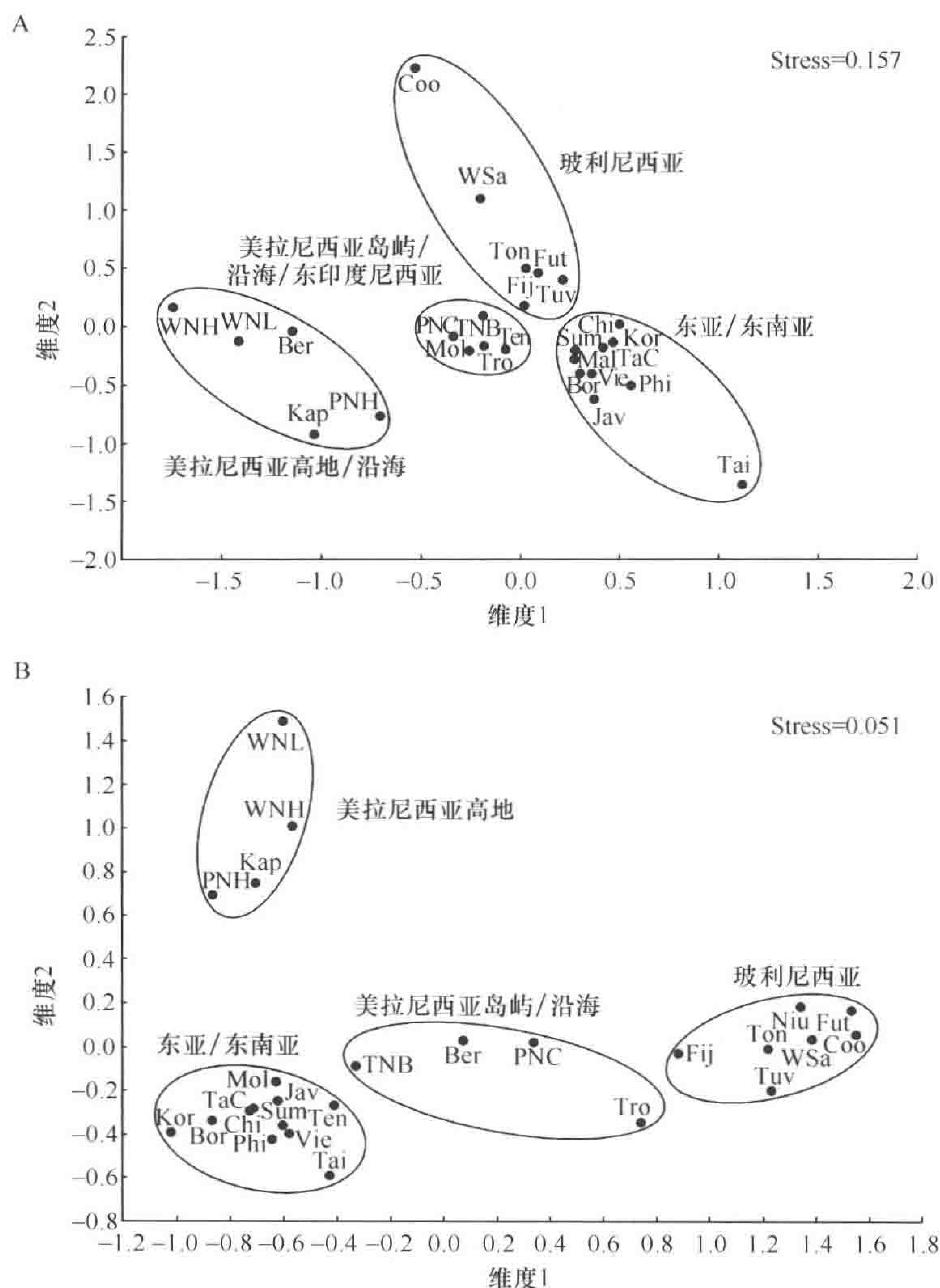



图 33-4　使用根据（A）NRY 单倍组和（B）mtDNA 的 HVl 序列资料进行的配对的群体比较得到的 Fst 值进行多维标度分析得到的二维地区图。群体的缩写方式与图 33-2 相同（一些样品较小的群体被省略掉了）。图中标明了群体样品的地理区域（Kayser et al.　2006）

在基于世系的分析时，单个世系是分析的单位，关注点是不同的世系之间怎样相互关联。通常一个世系的起源时间用序列的变异量（对于 mtDNA 世系）或 Y-STR 单倍型变异量（对于 Y-SNP 单倍组）来衡量。然而尽管基于世系的分析富有信息量，但要

记住为了获得群体历史的内情必须做额外的推论。例如，世系的年龄与群体的年龄不是一回事。而且因为当人们移动时，估计有许多世系一起移动，由此显示出迁徙的特征，很少能做到一个单一世系就标明一个迁徙过程。

对玻利尼西亚人资料的基于世系的分析，如利用对 NRY 单倍组用相关的 Y-STR 单倍型多样性或对 mtDNA 单倍组用 HVl 序列多样性所做的对每个单倍组分开进行的网状分析，揭示了主要的玻利尼西亚人单倍组具有与基于群体的分析相似的图画（图 33-5）。然而不同地理区域的差别不如基于群体的分析那样明显（除了 C-M208，C-M208 显示出玻利尼西亚人和美拉尼西亚人之间单倍型的完全分开）。而且网络结构图阐明了一个要点，即仅仅是这种频率的信息可能对于单倍组的起源是一种误导。这可以在 mtDNA 单倍组 B4a（图 33-5C）的网络结构图上看到：虽然在新几内亚 B4a 频率高

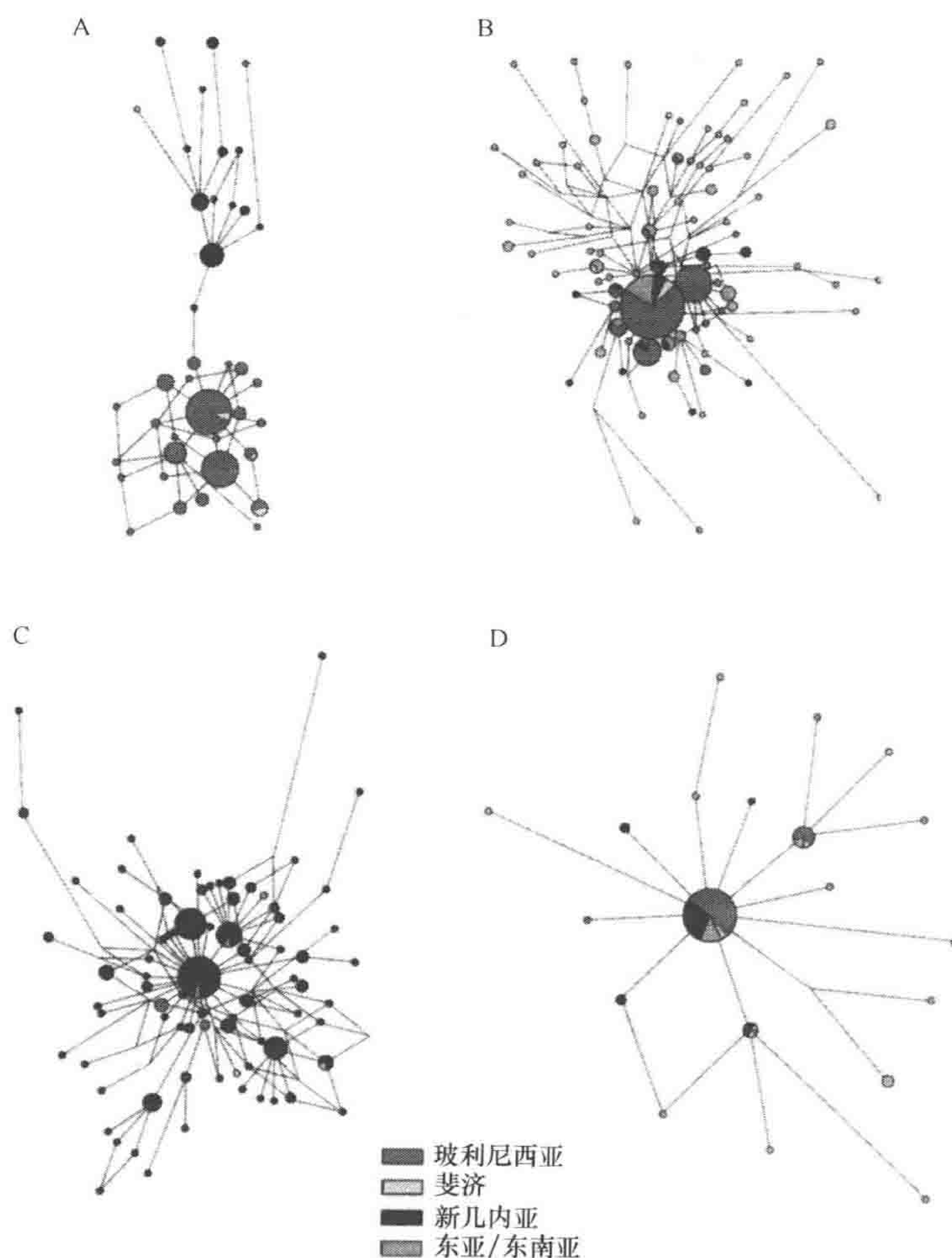



图 33-5　根据 7 个 Y-STR 位点，结合 NRY 单倍组（A）C-M208 和（B）O-M122 与 mtDNA单倍组关联的 HVl 序列（C）B4a 和（D）Q1，得到的单倍型的中点连接法网络结构图。圆圈表示单倍型，其面积与携带特定单倍型的个体数目成比例。线条表示突变的步骤。图中标明了个体样品的地理区域（Kayser et al.　2006）

于亚洲，这可以提示 B4a 的新几内亚起源，但与这个单倍组相关的 HVl 序列的多样性在亚洲最高（图 33-3），显示出这一单倍组起源于亚洲，而在新几内亚通过局部扩张提高了频率。进一步，所有在玻利尼西亚常见的 NRY 和 mtDNA 单倍组（C-M208、O-M122、B4a、PM 和 K-M9，这里没有显示最后两个）显示一种单倍型的网络结构，这种结构提示在玻利尼西亚人的遗传历史中出现的瓶颈事件，不同岛屿的许多玻利尼西亚人共享的一个或两个中央单倍型（也在玻利尼西亚人和亚洲/美拉尼西亚之间共享）。玻利尼西亚人在历史上经历的瓶颈在遗传学上有了强有力的证明。

关于玻利尼西亚人起源的最极端的模型（“快车”和“杂乱河堤”）并没有被 NRY 和 mtDNA 资料的综合分析所支持，这些资料显示亚洲人和美拉尼西亚人对玻利尼西亚人起源都有贡献。然而这些材料与我们多年前根据更有限的玻利尼西亚人的资料提出的“慢船”模型（“Slow boat” model，Kayser et al.　2000a）（以及其他居间模型，如“走出台湾”模型，这一模型没有指定玻利尼西亚祖先和美拉尼西亚人混合的类型和数量）相吻合。根据这一模型，玻利尼西亚人的祖先确实从东亚起源，但没有快速通过美拉尼西亚，他们与美拉尼西亚人交流和广泛混合，在定居太平洋之前留下了他们的基因，也整合了美拉尼西亚人的基因。而且，这种混合是高度不对等的，美拉尼西亚人主要通过男性混合入玻利尼西亚人的祖先，像美拉尼西亚人的 Y 染色体比他们的 mtDNA 对玻利尼西亚人的贡献更高所反映的那样，而玻利尼西亚人主要是女性祖先混合入美拉尼西亚人，像亚洲的 mtDNA 比他们的 Y 染色体对新几内亚人的贡献更高所反映的那样（Kayser et al.　2006）。有趣的是，玻利尼西亚的语言与玻利尼西亚的母性历史相吻合：玻利尼西亚人讲的太平洋中南诸岛的语言（Austronesian languages）很可能起源于中国台湾（Blust　1999）。近期另一个涉及南里海群体的 mtDNA 和 NRY 之间关系的差异的研究（Nasidze et al.　2006）也发现群体的语言与 mtDNA 的遗传相随，而不与 NRY 相随（在这个例子中语言和 mtDNA 都来自当地，而不是移动群体；而在玻利尼西亚的例子中 mtDNA 和语言来自移动群体，而不是当地的群体）。这种现象到底有多普遍还有待观察，但在任何情况下这些事例都证明了 mtDNA 和 NRY 的比较研究在洞察人类群体不同文化的父系和母系历史方面的独特能力。

## mtDNA 和 NRY 变异用于法律证明和系谱研究

如同上述讨论的，由于它们的特殊性质，特别是它们的单亲遗传类型，mtDNA 和 NRY 标记也用在法律证明和系谱研究中，以分别作为母系或父系世系的鉴定。在系谱研究中，它们适合于追溯个人感兴趣的父系或母系祖先。由于姓氏通常是通过父系向后代传递，大多数人类社会是父系世系，所以在系谱研究中，NRY 的遗传证据是结合姓氏信息以追踪男性个体的祖先。特别地，Y-STR 在系谱研究中十分流行，作为一种商业服务，一系列公司以系谱追踪为目的提供 Y-STR 类型。然而突变事件，加上隐藏的非生物学的父亲（以及较稀少的非生物学的母亲）可引起从家族记录和遗传资料得来的系谱信息的冲突，特别是对于倒回许多世代的系谱（Kayser et al.　2007）。对于最广泛使用的 Y-STR，其突变发生率高达每个单个标记每一千代平均 2～3 个突变（Kayser et

al. 2000b; Dupuy et al. 2004; Gusmao et al. 2005)。一系列基于DNA的家庭研究进行的人类鉴定的著名例子使用了当代的个体或与远古DNA分析相结合，都证明了mtDNA和NRY标记在系谱研究中的能量。这些著名例子包括对俄国沙皇尼古拉二世和他的家族骨骼的鉴定（Gill et al. 1994; Ivanov et al. 1996）和对托马斯·杰弗逊和艾斯顿·黑明斯·杰弗逊（杰弗逊的女奴莎莉·黑明斯的儿子）可能的父子关系的鉴定（Foster et al. 1998）。

以鉴定人为目的的法律证明在通常情况下使用常染色体的STR，在分析了一定数目的STR（一般10～15个）后即可提供基本的个体鉴定。已经建立了很大的STR罪犯资料库以备提供嫌疑人的信息。然而，在一些特殊的例子中常染色体STR提供的信息不足，而mtDNA和NRY标记则可以显示有用的信息。如在强奸案中NRY标记（Y-STR）被用来做男性系谱的鉴定。当女性受害者的DNA和男性犯罪者的DNA在样品中混合时可造成对男性施暴者常染色体STR分析的困难。而Y-STR甚至在精子不存在时（如涉及少精或无精男性的情况）可通过男性上皮细胞而被成功地探测到（Betz et al. 2001）。在多个男性犯罪者实施的性攻击时常染色体STR分析可能很成问题，但这种情况可以通过Y-STR解决，Y-STR可以区分不同的男性世系。NRY标记在根据男人的同父系的关系（并且因此有同一的或几乎同一的Y-STR构成）用DNA技术筛查寻找未确定的嫌疑人时也比常染色体更可取（Dettlaff-Kakol and Pawlowski 2002）。NRY标记也用于男性后代的父权确定，当推断的父亲无法做DNA鉴定时尤其如此。如果一个已故的推测父亲的双亲都不能做DNA鉴定，父子关系通常无法用常染色体STR非常确切地鉴定，但却可通过任何一个已故推测父亲的男性生物学亲属的Y-STR进行确定。

多拷贝的Y-STR具有多样的多态Y-STR位点，由于它们通常显示高度的多样性价值，曾经认为多拷贝的Y-STR（如DYS385a/b）在法律证明中应该特别有用。然而多拷贝的Y-STR应避免在罪犯的数目成问题时使用，如在轮奸案中（Butler et al. 2005）。需要格外注意的是一些多拷贝的Y-STR定位在与生育问题有关的重复的Y染色体区域（Bosch and Jobling 2003），因此它们的组成可以提供男性生育状态的信息，这是一个讨厌的信息（并在一些国家非法）。

由于mtDNA的多拷贝性，mtDNA通常是一些遗骸（如骨头或烧焦的尸体）中唯一的人类DNA鉴定的来源，这在大灾难发生时特别常见（Budowle et al. 2005）。如上所述，mtDNA也用于确定系谱中的母子关系。由于许多mtDNA和NRY单倍组有局限的地域分布，最近mtDNA和NRY标记又用来追踪样品提供者的地理起源或遗传的祖先。特别地，这样的信息可以在追踪疑犯而没有线索时帮助执法机构（Jobling and Gill 2004）。但为了后一种目的，mtDNA和NRY标记应该结合家系中有信息量的常染色体标记（Lao et al. 2006），以避免在个体有性别偏差的遗传混合历史时的单亲遗传信息的误读。

最后应该强调在非排除法律证明中，NRY和mtDNA标记用于犯罪者所属的父系或母系家系的鉴定，而不是真正的个体鉴定（主要从常染色体STR得到的鉴定）。由于mtDNA和NRY-DNA的非重组和单亲遗传，所有给定个体的母方亲属携带相同的

mtDNA 基因组，而所有给定男性父方的男性亲属携带相同的 Y 染色体（NRY）。因此举例来说，如果因为用常染色体标记做法律证明时不够成功而采用 NRY 和 mtDNA 分析，为确定对应于任何的母系的（mtDNA 证据）或父系的（NRY 证据）亲属嫌疑人是否犯了罪，进一步的调查是必须的。

## 致谢

我们感谢 Richard Cordaux 在网络方面的帮助、Knut Finstermeir 在画图方面的帮助，以及对本章所述研究作出贡献的我们的所有学生和同事们。

## 参考文献

Anderson S., Bankier A.T., Barrell B.G., de Bruijn M.H.L., Coulson, A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., et al. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290:** 457–465.

Awadalla P., Eyre-Walker A., and Smith J.M. 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science* **286:** 2524–2525.

Bamshad M.J., Watkins W.S., Dixon M.E., Jorde L.B., Rao B.B., Naidu J.M., Prasad B.V.R., Rasanayagam A., and Hammer M.F. 1998. Female gene flow stratifies Hindu castes. *Nature* **395:** 651–652.

Bandelt H.J., Quintana-Murci L., Salas A., and Macaulay V. 2002. The fingerprint of phantom mutations in mitochondrial DNA data. *Am. J. Hum. Genet.* **71:** 1150–1160.

Barbujani G., Vernesi C., Caramelli D., Castri L., Lalueza-Fox C., and Bertorelle G. 2004. Etruscan artifacts: Much ado about nothing. *Am. J. Hum. Genet.* **75:** 923–927.

Bellwood P. 2004. Colin Renfrew's emerging synthesis: Farming, languages and genes as viewed from the Antipodes. In *Traces of ancestry: Studies in honour of Colin Renfrew* (ed. M. Jones), pp. 31–39. McDonald Institute, Cambridge, United Kingdom.

Bendall K.E. and Sykes B.C. 1995. Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am. J. Hum. Genet.* **57:** 248–256.

Betz A., Bassler G., Dietl G., Steil X., Weyermann G., and Pflug W. 2001. DYS STR analysis with epithelial cells in a rape case. *Forensic Sci. Int.* **118:** 126–130.

Blust R. 1999. Subgrouping, circularity and extinction: Some issues in Austronesian comparative linguistics. *Symp. Ser. Inst. Linguistics Acad. Sinica* **1:** 31–94.

Bosch E. and Jobling M.A. 2003. Duplications of the *AZFa* region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. *Hum. Mol. Genet.* **12:** 341–347.

Budowle B., Bieber F.R., and Eisenberg A.J. 2005. Forensic aspects of mass disasters: Strategic considerations for DNA-based human identification. *Leg. Med.* **7:** 230–243.

Butler J.M., Decker, A.E., Kline, M.C., and Vallone, P.M. 2005. Chromosomal duplications along the Y-chromosome and their potential impact on Y-STR interpretation. *J. Forensic Sci.* **50:** 853–859.

Cann R.L., Stoneking M., and Wilson A.C. 1987. Mitochondrial DNA and human evolution. *Nature* **325:** 31–36.

Casanova M., Leroy P., Boucekkine C., Weissenbach J., Bishop C., Fellous M., Purrello M., Fiori G., and Siniscalco M. 1985. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science* **230:** 1403–1406.

de Knijff P. 2000. Messages through bottlenecks: On the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am. J. Hum. Genet.* **67:** 1055–1061.

Dettlaff-Kakol A. and Pawlowski R. 2002. First Polish DNA "manhunt"—An application of Y-chromosome STRs. *Int. J. Leg. Med.* **116:** 289–291.

Diamond J. 1988. Express train to Polynesia. *Nature* **336:** 307–308.

Dupuy B.M., Stenersen M., Egeland T., and Olaisen B. 2004. Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. *Hum. Mutat.* **23:** 117–124.

Excoffier L. 1990. Evolution of human mitochondrial DNA: Evidence for departure from a pure neutral model of populations at equilibrium. *J. Mol. Evol.* **30:** 125–139.

Forster P. 2003. To err is human. *Ann. Hum. Genet.* **67:** 2–4.

Foster E.A., Jobling M.A., Taylor P.G., Donnelly P., de Knijff P., Mieremet R., Zerjal T., and Tyler-Smith C. 1998. Jefferson fathered slave's last child. *Nature* **396:** 27–28.

Giles R.E., Blanc H., Cann H.M., and Wallace D.C. 1980. Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci.* **77:** 6715–6719.

Gill P., Ivanov P.L., Kimpton C., Piercy R., Benson N., Tully G., Evett I., Hagelberg E., and Sullivan K. 1994. Identification of the remains of the Romanov family by DNA analysis. *Nat. Genet.* **6:** 130–135.

Green R.C. 1991. The Lapita cultural complex: Current evidence and proposed models. *Bull. Indo-Pacific Prehist. Assoc.* **11:** 295–305.

Green R.E., Krause J., Ptak S.E., Briggs A.W., Ronan M.T., Simons J.F., Du L., Egholm M., Rothberg J.M., Paunovic M., and Paabo S. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444:** 330–336.

Groube L.M., Chappell J., Muke J., and Price D. 1986. A 40,000 year-old human occupation site at Huon Peninsula, Papua New Guinea. *Nature* **324:** 453–455.

Gusmao L., Sanchez-Diz P., Calafell F., Martin P., Alonso C.A., Alvarez-Fernandez F., Alves C., Borjas-Fajardo L., Bozzo W.R., Bravo M.L., et al. 2005. Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.* **26:** 520–528.

Hage P. 1998. Was Proto-Oceanic society matrilineal? *J. Polynesian Soc.* **107:** 365–379.

Hagelberg E., Goldman N., Liò P., Whelan S., Schiefenhövel W., Clegg J.B., and Bowden D.K. 1999. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proc. R. Soc. Lond. B Biol. Sci.* **266:** 485–492.

———. 2000. Evidence for mitochondrial DNA recombination

in a human population of island Melanesia: Correction. *Proc. R. Soc. Lond. B Biol. Sci.* **267:** 1595–1596.

Ingman M., Kaessmann H., Paabo S., and Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* **408:** 708–713.

Ivanov P.L., Wadhams M.J., Roby R.K., Holland M.M., Weedn V.W., and Parsons T.J. 1996. Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nat. Genet.* **12:** 417–420.

Jobling M.A. and Gill P. 2004. Encoded evidence: DNA in forensic analysis. *Nat. Rev. Genet.* **5:** 739–751.

Jobling M.A. and Tyler-Smith C. 2003. The human Y chromosome: An evolutionary marker comes of age. *Nat. Rev. Genet.* **4:** 598–612.

Johnson M.J., Wallace D.C., Ferris S.D., Rattazzi M.C., and Cavalli-Sforza L.L. 1983. Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. *J. Mol. Evol.* **19:** 255–271.

Kayser M., Vermeulen M., Knoblauch H., Schuster H., Krawczak M., and Roewer L. 2007. Relating two deep-rooted pedigrees from Central Germany by high-resolution Y-STR haplotyping. *Forensic Sci. Int. Genet.* **1:** 125–128.

Kayser M., Brauer S., Weiss G., Underhill P.A., Roewer L., Schiefenhovel W., and Stoneking M. 2000a. Melanesian origin of Polynesian Y chromosomes. *Curr. Biol.* **10:** 1237–1246.

Kayser M., Brauer S., Weiss G., Schiefenhovel W., Underhill P., Shen P., Oefner P., Tommaseo-Ponzetta M., and Stoneking M. 2003. Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am. J. Hum. Genet.* **72:** 281–302.

Kayser M., Brauer S., Cordaux R., Casto A., Lao O., Zhivotovsky L.A., Moyse-Faurie C., Rutledge R.B., Schiefenhoevel W., Gil D., et al. 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol. Biol. Evol.* **23:** 2234–2244.

Kayser M., Caglia A., Corach D., Fretwell N., Gehrig C., Graziosi G., Heidorn F., Herrmann S., Herzog B., Hidding M., et al. 1997. Evaluation of Y-chromosomal STRs: A multicenter study. *Int. J. Leg. Med.* **110:** 125–133, 141–149.

Kayser M., Kittler R., Erler A., Hedman M., Lee A.C., Mohyuddin A., Mehdi S.Q., Rosser Z., Stoneking M., Jobling M.A., et al. 2004. A comprehensive survey of human Y-chromosomal microsatellites. *Am. J. Hum. Genet.* **74:** 1183–1197.

Kayser M., Krawczak M., Excoffier L., Dieltjes P., Corach D., Pascali V., Gehrig C., Bernini L.F., Jespersen J., Bakker E., et al. 2001. An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am. J. Hum. Genet.* **68:** 990–1018.

Kayser M., Lao O., Anslinger K., Augustin C., Bargel G., Edelmann J., Elias S., Heinrich M., Henke J., Henke L., et al. 2005. Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. *Hum. Genet.* **117:** 428–443.

Kayser M., Roewer L., Hedman M., Henke L., Henke J., Brauer S., Kruger C., Krawczak M., Nagy M., Dobosz T., et al. 2000b. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* **66:** 1580–1588.

Kirch P.V. 2000. *On the road of the winds: An archaeological history of the Pacific Islands before European contact.* University of California Press, Berkeley.

Kittler R., Erler A., Brauer S., Stoneking M., and Kayser M. 2003. Apparent intrachromosomal exchange on the human Y chromosome explained by population history. *Eur. J. Hum. Genet.* **11:** 304–314.

Kivisild T., Shen P., Wall D.P., Do B., Sung R., Davis K., Passarino G., Underhill P.A., Scharfe C., Torroni A., et al. 2006. The role of selection in the evolution of human mitochondrial genomes. *Genetics* **172:** 373–387.

Kraytsberg Y., Schwartz M., Brown T.A., Ebralidse K., Kunz W.S., Clayton D.A., Vissing J., and Khrapko K. 2004. Recombination of human mitochondrial DNA. *Science* **304:** 981.

Krings M., Stone A., Schmitz R.W., Krainitzki H., Stoneking M., and Paabo S. 1997. Neandertal DNA sequences and the origin of modern humans. *Cell* **90:** 19–30.

Kumar S., Hedrick P., Dowling T., and Stoneking M. 2000. Questioning evidence for recombination in human mitochondrial DNA. *Science* **288:** 1931.

Lao O., van Duijn K., Kersbergen P., de Knijff P., and Kayser M. 2006. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am. J. Hum. Genet.* **78:** 680–690.

Macpherson J.M., Ramachandran S., Diamond L., and Feldman M.W. 2004. Demographic estimates from Y chromosome microsatellite polymorphisms: Analysis of a worldwide sample. *Hum. Genomics* **1:** 345–354.

Mishmar D., Ruiz-Pesini E., Golik P., Macaulay V., Clark A.G., Hosseini S., Brandon M., Easley K., Chen E., Brown M.D., et al. 2003. Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci.* **100:** 171–176.

Nachman M., Brown W., Stoneking M., and Aquadro C. 1996. Non-neutral mitochondrial DNA variation in humans and chimpanzees. *Genetics* **142:** 953–963.

Nasidze I., Quinque D., Rahmani M., Alemohamad S.A., and Stoneking M. 2006. Concomitant replacement of language and mtDNA in South Caspian populations of Iran. *Curr. Biol.* **16:** 668–673.

Nasidze I., Quinque D., Dupanloup I., Cordaux R., Kokshunova L., and Stoneking M. 2005. Genetic evidence for the Mongolian ancestry of Kalmyks. *Am. J. Phys. Anthropol.* **128:** 846–854.

Nasidze I., Ling E.Y., Quinque D., Dupanloup I., Cordaux R., Rychkov S., Naumova O., Zhukova O., Sarraf-Zadegan N., Naderi G.A., et al. 2004. Mitochondrial DNA and Y-chromosome variation in the Caucasus. *Ann. Hum. Genet.* **68:** 205–221.

Noonan J.P., Coop G., Kudaravalli S., Smith D., Krause J., Alessi J., Chen F., Platt D., Paabo S., Pritchard J.K., and Rubin E.M. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* **314:** 1113–1118.

Oota H., Settheetham-Ishida W., Tiwawech D., Ishida T., and Stoneking M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nat. Genet.* **29:** 20–21.

Pakendorf B., Novgorodov I.N., Osakovskij V.L., Danilova A.P., Protod'jakonov A.P., and Stoneking M. 2006. Investigating the effects of prehistoric migrations in Siberia: Genetic variation and the origins of Yakuts. *Hum. Genet.* **120:** 334–353.

Pakendorf B. and Stoneking M. 2005. Mitochondrial DNA and human evolution. *Annu. Rev. Genomics Hum. Genet.* **6:** 165–183.

Piganeau G. and Eyre-Walker A. 2004. A reanalysis of the indirect evidence for recombination in human mitochondrial DNA. *Heredity* **92:** 282–288.

Redd A.J., Agellon A.B., Kearney V.A., Contreras V.A., Karafet T., Park H., de Knijff P., Butler J.M., and Hammer M.F. 2002. Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci. Int.* **130:** 97–111.

Roberts R.G., Jones R., and Smith M.A. 1990. Thermoluminescence dating of a 50,000-year-old human occupation site in northern Australia. *Nature* **345:** 153–156.

Roewer L., Arnemann J., Spurr N.K., Grzeschik K.-H., and Epplen J.T. 1992. Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Hum. Genet.* **89:** 389–394.

Roewer L., Croucher P.J., Willuweit S., Lu T.T., Kayser M., Lessig R., de Knijff P., Jobling M.A., Tyler-Smith C., and Krawczak M. 2005. Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Hum. Genet.*

**116**: 279–291.

Rogers A.R. and Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9:** 552–569.

Rosser Z.H., Zerjal T., Hurles M.E., Adojaan M., Alavantic D., Amorim A., Amos W., Armenteros M., Arroyo E., Barbujani G., et al. 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.* **67:** 1526–1543.

Ruiz-Pesini E., Mishmar D., Brandon M., Procaccio V., and Wallace D.C. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303:** 223–226.

Salas A., Carracedo A., Macaulay V., Richards M., and Bandelt H.J. 2005. A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. *Biochem. Biophys. Res. Commun.* **335:** 891–899.

Schwartz M. and Vissing J. 2002. Paternal inheritance of mitochondrial DNA. *N. Engl. J. Med.* **347:** 576–580.

Seielstad M., Minch E., and Cavalli-Sforza L. 1998. Genetic evidence for a higher female migration rate in humans. *Nat. Genet.* **20:** 278–280.

Seielstad M., Yuldasheva N., Singh N., Underhill P., Oefner P., Shen P., and Wells R.S. 2003. A novel Y-chromosome variant puts an upper limit on the timing of first entry into the Americas. *Am. J. Hum. Genet.* **73:** 700–705.

Semino O., Passarino G., Oefner P.J., Lin A.A., Arbuzova S., Beckman L.E., De Benedictis G., Francalacci P., Kouvatsi A., Limborska S., et al. 2000. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: A Y chromosome perspective. *Science* **290:** 1155–1159.

Serre D., Langaney A., Chech M., Teschler-Nicola M., Paunovic M., Mennecier P., Hofreiter M., Possnert G., and Paabo S. 2004. No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol.* **2:** E57.

Skaletsky H., Kuroda-Kawaguchi T., Minx P.J., Cordum H.S., Hillier L., Brown L.G., Repping S., Pyntikova T., Ali J., Bieri T., et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825–837.

Sun C., Kong Q.P., and Zhang Y.P. 2006. The role of climate in human mitochondrial DNA evolution: A reappraisal. *Genomics* **89:** 338–342.

Terrell J.E. 1988. History as a family tree, history as an entangled bank: Constructing images and interpretations of prehistory in the South Pacific. *Antiquity* **62:** 642–657.

Thangaraj K., Chaubey G., Kivisild T., Reddy A.G., Singh V.K., Rasalkar A.A., and Singh L. 2005. Reconstructing the origin of Andaman Islanders. *Science* **308:** 996.

Thomson R., Pritchard J.K., Shen P., Oefner P.J., and Feldman M.W. 2000. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc. Natl. Acad. Sci.* **97:** 7360–7365.

Torroni A., Schurr T.G., Yang C.-C., Szathmary E.J.E., Williams R.C., Schanfield M.S., Troup G.A., Knowler W.C., Lawrence D.N., Weiss K.M., and Wallace D.C. 1992. Native American mitochondrial DNA analysis indicates that the Amerind and Nadene populations were founded by two independent migrations. *Genetics* **130:** 153–162.

Trejaut J.A., Kivisild T., Loo J.H., Lee C.L., He C.L., Hsu C.J., Lee Z.Y., and Lin M. 2005. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biol.* **3:** e247.

Underhill P.A., Shen P., Lin A.A., Jin L., Passarino G., Yang W.H., Kauffman E., Bonne-Tamir B., Bertranpetit J., Francalacci P., et al. 2000. Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26:** 358–361.

Vigilant L., Stoneking M., Harpending H., Hawkes K., and Wilson A.C. 1991. African populations and the evolution of human mitochondrial DNA. *Science* **253:** 1503–1507.

Wall J.D. and Hammer M.F. 2006. Archaic admixture in the human genome. *Curr. Opin. Genet. Dev.* **16:** 606–610.

Wilder J.A., Kingan S.B., Mobasher Z., Pilkington M.M., and Hammer M.F. 2004. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nat. Genet.* **36:** 1122–1125.

Y Chromosome Consortium. 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12:** 339–348.

Zerjal T., Dashnyam B., Pandya A., Kayser M., Roewer L., Santos F.R., Schiefenhovel W., Fretwell N., Jobling M.A., Harihara S., et al. 1997. Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. *Am. J. Hum. Genet.* **60:** 1174–1183.

Zerjal T., Xue Y., Bertorelle G., Wells R.S., Bao W., Zhu S., Qamar R., Ayub Q., Mohyuddin A., Fu S., et al. 2003. The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72:** 717–721.

# 34 适于法庭的 DNA 测试

John M. Butler

*Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8311*



## 简介

DNA 测试和遗传变异的检测最广泛的应用领域之一是对人的鉴定，这归功于像 *CSI*（犯罪现场调查）这类电视节目的普及。在好莱坞的噱头之外，在现实中适于法庭的 DNA 分型已经解决了无数的犯罪问题——既有提示有罪的，也有证明无辜的。

这些技术和遗传标记还有许多其他的应用范围。在美国每年有成百万的样品用于亲子鉴定，主要是为孩子的抚养确定父亲身份（AABB 2005）。在美国军队中DNA测试用于确定战争伤亡人员，以避免下一个“无名士兵”。自然的或人为的灾难或恐怖主义袭击中受害人的遗留物可用生物学亲属或直接相关样品的DNA测试加以确认（Whitaker et al. 1995；Biesecker et al. 2005；NIJ 2006）。保存了成百万的罪犯的遗传学文件的国家DNA数据库被美国、英国及许多其他国家用于寻找作案者，这在几年前还是不可能的（Gill 2002）。

在试图解决历史遗留问题，如托马斯·杰弗逊是否是一个奴隶的父亲（Foster et al. 1998），或鉴定罗曼诺夫家族的遗骨（Gill et al. 1994）时，同样运用了法庭DNA测试技术。此外遗传系谱学已经变成了重新建立模糊不清的家族系谱的流行工具（Brown 2002）。

在过去20年中，在分子水平上检测遗传变异的技术飞速发展。Alec Jeffrey在1984年发现的小卫星DNA和高变的串联重复（Jeffrey et al. 1985）被认为是现代法律DNA检测的开始，不过多位点的可变数目串联重复序列（VNTR）探针技术（Gill et al. 1985）在近10年已被短串联重复序列（STR）分型所取代（NIJ 2000；Butler 2005）。

STR有时称为微卫星DNA或简单序列重复（SSR），是一段手风琴样的DNA片段，包含有2～7个核苷酸的核心重复单元串联排列，一般重复了6次到几十次（图34-1）。虽然人类基因组有成百万的STR标记，但只有一小部分位点被用来做法律DNA检测（Butler 2006）。商业的试剂盒（表34-1）已被用于检测这些STR位点（表




图34-1　A. 表示一个STR标记的序列的例子，下划线核苷酸序列表示PCR引物结合位点，粗体字区域显示12个GATA四核苷酸重复。B. 表示有7～13次重复的可能的等位基因的长度，箭头表示PCR引物的位置

34-2)。每年政府、大学或私人实验室都进行成百万的这类遗传变异检测，用于 DNA 数据库建设、法律案件或亲子鉴定。

**表 34-1　一些广泛应用的常染色体 STR 标记的商业试剂盒**

| 试剂盒名称 | 来源 | 包括的 STR 位点<br>Amel＝性分型标记 amelogenin | 辨别能力[a] |
|---|---|---|---|
| PowerPlex16 | Promega | CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, amel, D16S539, D18S51, D21S11, Penta D, Penta E | 1 in $3.4\times10^{17}$ |
| Profiler Plus | Applied Biosystems | FGA, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, amel | 1 in $9.4\times10^{10}$ |
| COfiler | Applied Biosystems | CSF1PO, TH01, TPOX, D3S1358, D7S820, D16S539, amel | 1 in $1.3\times10^{6}$ |
| SGM Plus | Applied Biosystems | FGA, TH01, VWA, D3S1358, D8S1179, D16S539, D18S51, D21S11, D2S1338, D19S433, amel | 1 in $4.8\times10^{12}$ |
| Identifiler | Applied Biosystems | CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, D2S1338, D19S433, amel | 1 in $2.5\times10^{17}$ |

a 依据在表 34-2 中的每一个 STR 位点随机配对可能性。

**表 34-2　常见的常染色体 STR 标记特征**

| STR 标记 | 染色体位置 | 重复的片段 | 等位基因范围[a] | PCR 产物的大小 | RMP[b] |
|---|---|---|---|---|---|
| CSF1PO | 5q33.1 | TAGA | 6～15 | 305～342bp (identifiler) | 0.112 |
| FGA | 4q31.3 | CTTT | 17～51.2 | 215～355bp (identifiler) | 0.036 |
| TH01 | 11p15.5 | TCAT | 4～13.3 | 163～202bp (identifiler) | 0.081 |
| TPOX | 2p25.3 | GAAT | 6～13 | 222～250bp (identifiler) | 0.195 |
| VWA | 12p13.31 | [TCTG][TCTA] | 11～24 | 155～207bp (identifiler) | 0.062 |
| D3S1358 | 3p21.31 | [TCTG][TCTA] | 12～19 | 112～140bp (identifiler) | 0.075 |
| D5S818 | 5q23.2 | AGAT | 7～16 | 134～172bp (identifiler) | 0.158 |
| D7S820 | 7q21.11 | GATA | 6～15 | 255～291bp (identifiler) | 0.065 |
| D8S1179 | 8q24.13 | [TCTA][TCTG] | 8～19 | 123～170bp (identifiler) | 0.067 |
| D13S317 | 13q31.1 | TATC | 8～15 | 217～245bp (identifiler) | 0.085 |

续表

| STR 标记 | 染色体位置 | 重复的片段 | 等位基因范围[a] | PCR 产物的大小 | RMP[b] |
|---|---|---|---|---|---|
| D16S539 | 16q24.1 | GATA | 5～15 | 252～292bp (identifiler) | 0.089 |
| D18S51 | 18q21.33 | AGAA | 7～27 | 262～345bp (identifiler) | 0.028 |
| D21S11 | 21q21.1 | [TCTA][TCTG] | 24～38 | 185～239bp (identifiler) | 0.039 |
| D2S1338 | 2q35 | [TGCC][TTCC] | 15～28 | 307～359bp (identifiler) | 0.027 |
| D19S433 | 19q12 | AAGG | 9～17.2 | 102～135bp (identifiler) | 0.087 |
| Penta D | 21q22.3 | AAAGA | 2.2～17 | 376～449bp(PP16) | 0.059 |
| Penta E | 15q26.2 | AAAGA | 5～24 | 379～474bp(PP16) | 0.030 |
| Amelogenin (sex-typing) | Xp22.22 Yp11.2 | 不适用 | | X=107bp Y=113bp | |

a 范围是通过试剂盒等位阶梯的计算得来，不代表在世界群体中观察到的等位基因的全部范围。（这些 STR 位点的更完整的等位基因列表可参见 Butler 2006）；b RMP：高加索个体随机配对可能性（Random match probability）。13 个 CODIS 位点资料来自 NIJ（2000）所报告的 FBI 资料。D2S1138 和 D19S433 RMP 资料来自 Applied Biosystems 的 Identifiler 试剂盒用户指南。Penta D 和 Penta E 的信息来自 Promega 公司的 PowerPlex 16（PP16）群体资料。

## 法律 DNA 检测的总体情况

由于在减数分裂时的重组和染色体分离，每一个个体遗传自父母的基因组都是独特的，只有同卵双生是例外。然而由于检查个体完整基因组在价格和时间上的限制，只有少数亚群用来探测遗传变异，在人类鉴定和法律 DNA 检测中用于区分个体。因而根据 DNA 图谱上不同等位基因的期待频率使用随机配对的统计概率。

由于在 DNA 操作时收集的信息要用于法庭，要在分析过程的每一个步骤跟踪样品，并申明委托保管流程，此外，只能使用有效的检测方法以确保取得可靠和可重复的结果。目前的标准 DNA 测试不检测基因，而是检测发生在内含子或基因间的“垃圾”区域的变异，认识这一点很重要。因此标准的法庭 DNA 检测不能提供有关种族，或如眼睛颜色、身高、头发颜色等表型的信息。

STR 分型运用 PCR 从少量的材料中提取信息。在犯罪现场发现的作为证据的生物材料常由于环境因素导致 DNA 发生降解，而 STR 实验产生的长度为 100～500 bp 的较短 PCR 产物可从降解材料得到。使用不同颜色的荧光染料和不同大小的 PCR 产物，多 STR 位点同时进行 PCR 扩增，或使用“多路技术”是可以做到的。多位点的使用可以在单一试验中在不消耗许多 DNA 样品（如使用 1ng 或更少的起始材料）的情况下具有较高的辨别力。商用的试剂盒简化了 STR 的步骤，并且能提供核心 STR 位点的一致结果，这使国内和国际共享罪犯 DNA 资料成为可能。商用试剂盒较贵，但它使操作简单

化和标准化，并且免除了忙碌的使用者的 PCR 成分质量控制方面的负担，因而适合于自主的检验。STR 分型（图 34-2）的步骤包括样品采集、DNA 提取、DNA 定量、多 STR 位点的 PCR 扩增、STR 等位的分离和长短测量、STR 分型及其解释、匹配的统计学显著性分析报告（如果观察到了的话）等。在后面章节中对这些步骤有详细的描述（也可参见 Budowle et al. 2000）。

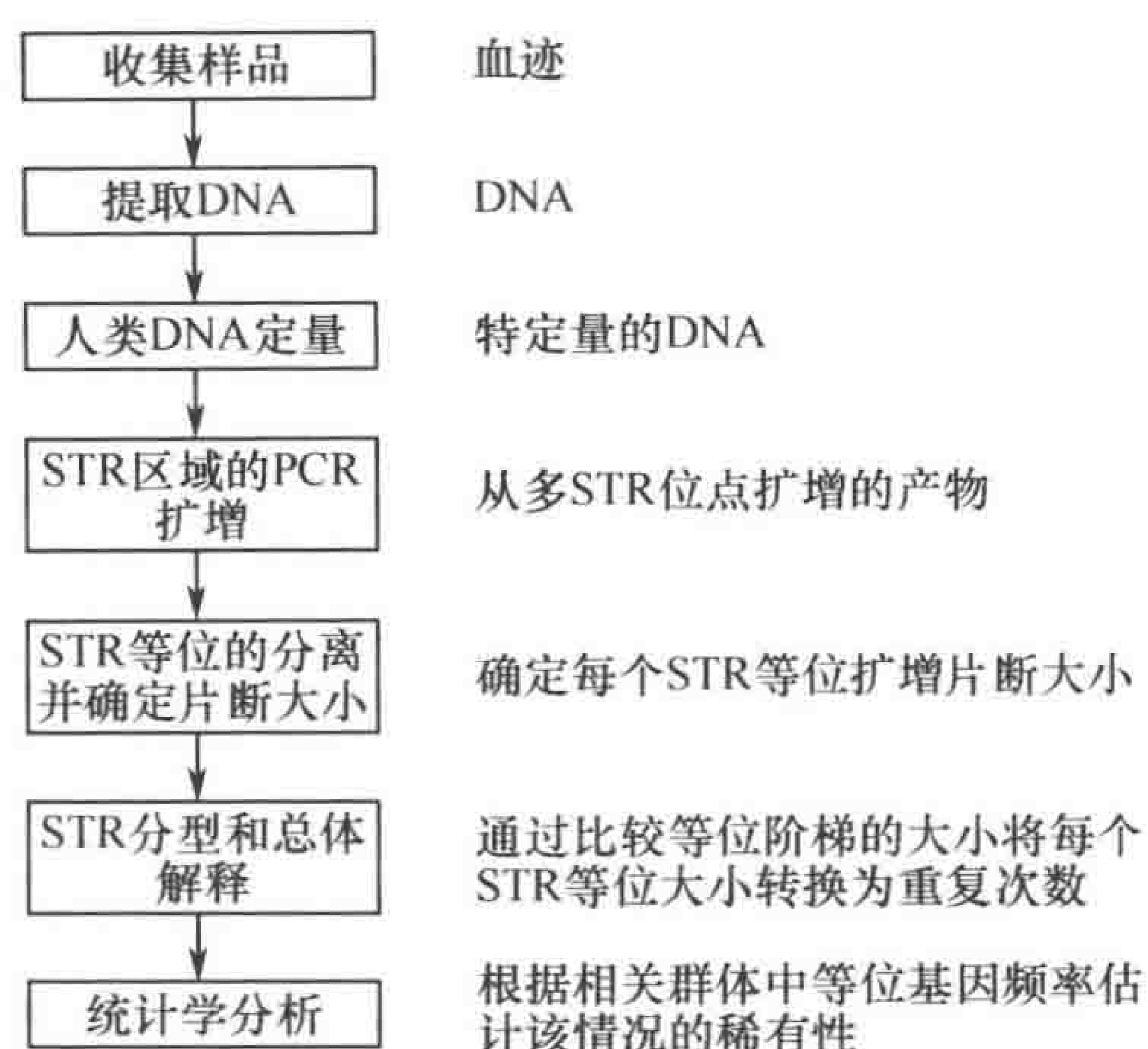



图 34-2　STR 分型各步骤和各步骤产出的总结

## 标本采集

涂抹在纸卡片的血液，或从个体口腔内刮取得颊黏膜细胞是从嫌疑人或其生物学亲属等相关样品搜集 DNA 的常用方法。犯罪现场的斑迹可来自剪下的衣物一角，在强奸案例中的受害者阴道试子等。

## DNA 提取

为了适合于 PCR 引物和其他试剂，DNA 必须从细胞中提取出来并形成溶液。法庭实验室适用的两种方便的 DNA 提取方法是有机提取和 Chelex 提取。有机提取涉及十二烷基硫酸钠和蛋白酶 K 的连续稀释以破坏细胞壁，随后加入酚-氯仿混合液以便将蛋白质和 DNA 分离。通过离心法将蛋白质和其他细胞碎片从水相中除去，而双链 DNA 留在水相中。样品常通过 Centricon 或 Microcon 透析加以浓缩（Comey et al. 1994）。

Chelex 使用螯合离子交换树脂悬液，将该悬液直接加到样品中（Walsh et al. 1991）。许多方案将血迹加到 5%的 Chelex 悬液中，煮沸几分钟破坏细胞并释放出 DNA，然后离心，取含有变性 DNA 分子的上清液进行提取后的步骤。

许多样品是在 FTA 纸（Whatman）上的干燥的血液或唾液，在纸上打孔得到的样品部分用于 PCR。在经过几次化学清洗以便除去 PCR 抑制复合物后，打孔得到的清洁的结合着 DNA 的 FAT 小片直接加入到 PCR 反应体系中（Vanek et al. 2001）。此外，许多基于唾液的固相提取方法的自动化程序，如 QIAamp spin columns（Qiagen）和 DNA IQ 磁珠（Promega）等，已经在法庭 DNA 界得到普遍使用。

在性袭击发生射精时，阴道试子含有女性受害者和男性罪犯的DNA混合样品。在性袭击取证时常采用一种称为分别提取（differential extraction）的步骤，先将精子和阴道内皮细胞进行物理分离，然后再破碎细胞，释放DNA（第3章，Butler 2005）。分别提取涉及移去女性上皮细胞后使用二硫苏糖醇（DTT）裂解精子细胞（Gill et al. 1985）。

## DNA定量

在从犯罪现场找到的生物学证据中提取DNA时，像细菌、植物或动物等非人类来源的DNA可能被同时提取，由于在PCR时只希望人类的DNA被扩增，应该确定得到的是特定人类DNA量。由于qPCR方法的敏感度、人类特异性和动态范围，在法律界内该方法在DNA定量时已经普遍使用。现在经常用Quantifiler试剂盒（Applied Biosystems）做qPCR检测。

使用STR试剂盒进行多元PCR扩增十分敏感，在很小的给定DNA量的范围内效果很好。加入DNA的精确定量有助于产生平衡的、稳定的STR结果。在高于2ng时，荧光检测系统信号过强，造成许多人造假相，使资料的解释变得困难。由于不完全的腺嘌呤化，或超出正常范围的峰值，造成不同染色之间的相互渗透，太多的DNA可形成劈裂的峰。在比200pg低很多的情况下，随机扩增可能导致位点或等位的失衡，甚至遗失。DNA浓度的调整通常取决于定量的结果，因此加入到PCR体系的DNA模板量应为0.5～1.5ng，可确保得到平衡的和稳定的STR结果。

## PCR扩增

通过提供先期混合的引物，标准化的酶缓冲液和dNTP的精确混合，商用STR试剂盒极大地简化了PCR扩增的步骤。使用试剂盒尽管昂贵，但免除了使用者对PCR反应成分进行质量控制的负担，特别是在多重扩增中有大于30个寡核苷酸引物存在时，如使用PowerPlex 16（Krenke et al. 2002）和Identifiler kits（Collins et al. 2004）时非常方便。AmpliTaq Gold DNA聚合酶（Applied Biosystems）通常被用来确保特异性的热启动反应。典型的扩增策略是在95℃下10min以激活TaqGold聚合酶，然后在94℃下变性，在59℃或60℃使引物退火，72℃延伸，共进行28～30个循环。

## STR等位的分离和确定片段长度

PCR扩增后，通过测量STR扩增片段的长度来确定DNA样品中每个等位的重复次数。这种长度的测量是通过基于片段大小的凝胶或毛细管电泳来实现的。因为正向和逆向的位点特异性引物都含有荧光染料，每个STR扩增子在PCR中都被荧光标记。这样通过记录每个DNA片段的染料颜色和相对于内部长度标准物的迁移时间，在每个STR等位与其他STR等位分离后，其长度都可以被确定下来。STR等位分离和片段长度确定使用的常用仪器包括ABI 310和ABI 3100 Genetic Analyzers（Buffer et al. 2004）。

## STR 分型和整体解读

商用 STR 试剂盒与等位阶梯一起提供，等位阶梯是通过 DNA 测序得到的以重复单元的数目为特征的一般等位的混合体。等位片段通过与等位阶梯作比较确定自己的大小（与同一片段内部标准物共同进行电泳），可将每个样品等位片段根据碱基对数所确定的大小转变为重复单元的数目（图 34-3）。等位阶梯允许个体的重复长度对其等位片段进行校准，这使来自不同电泳条件的实验室得到的数据之间的比较成为可能。对于不同的 STR 位点分析来说，最终的 DNA 资料都是基因型的线性结构。下面列举了两个例子，每一个都包含了从法庭 DNA 研究团体常使用的 13 个核心 STR 位点得来的基因型分型。

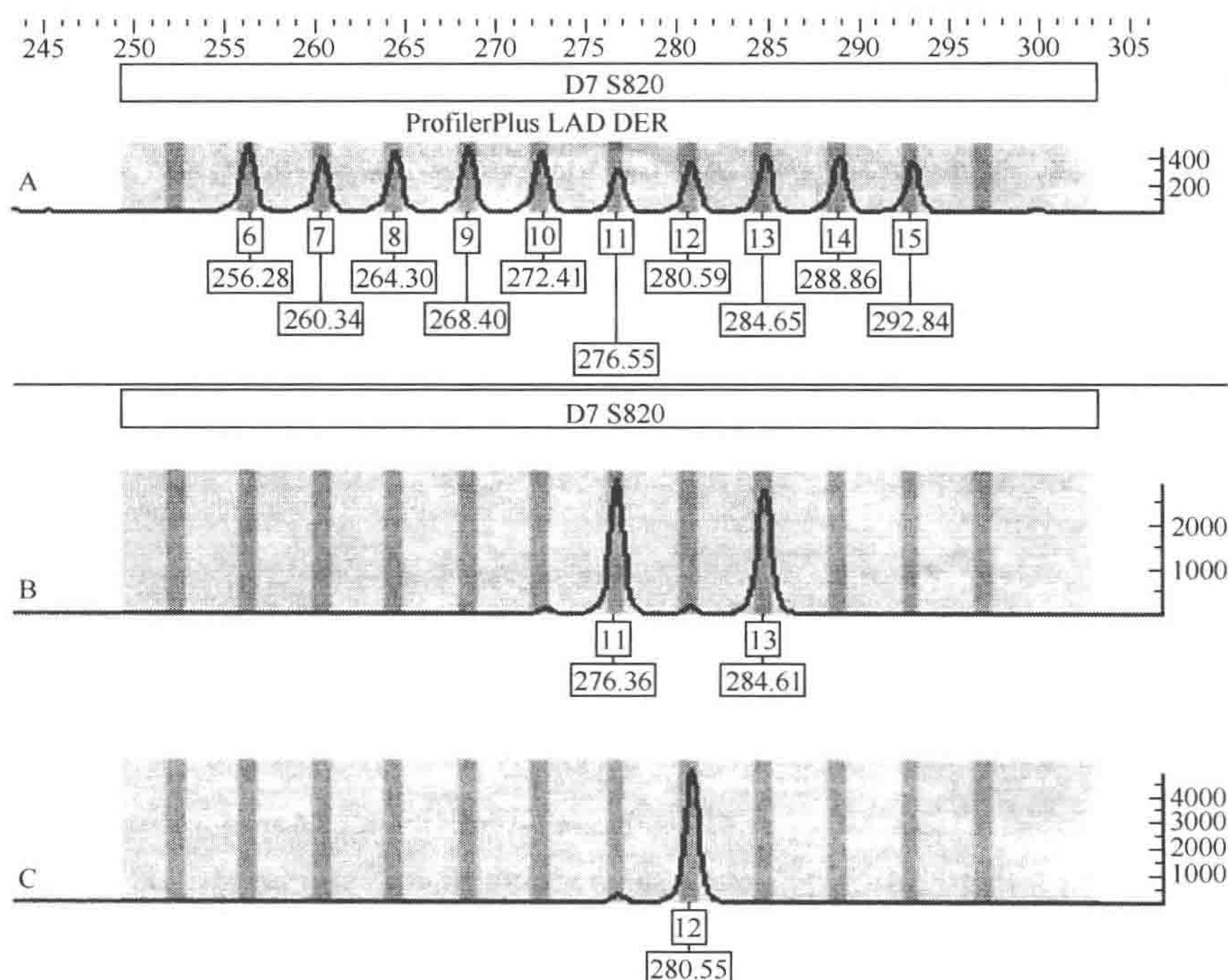



图 34-3　A. 表示与两个样品一起显示的 STR 位点 D7S820 的等位阶梯；B. 表示杂合子 11，13 和 C. 表示纯合子 12，12

第一组：16，17-17，18-21，22-12，14-28，30-14，16-12，13-11，14-9，9-9，11-6，6-8，8-10，10。

第二组：14，17-17，17-21，22-13，13-29，33.2-14，14-12，12-11，13-9，10-12，14-6，7-8，11-12，12。

这些组对于每个位点（由短杠分开）包含了两个等位（由逗号隔开）。这个例子中 STR 位点的顺序是 D3S1358，VWA，FGA，D8S1179，D21S11，D18S51，D5S818，D13S317，D7S820，D16S539，TH01，TPOX，CSFIPO（表 34-2）（对每个位点的更多的信息参见 Butler 2005，2006）。图 34-4 显示对第二组的电泳图谱（electropherogram）。

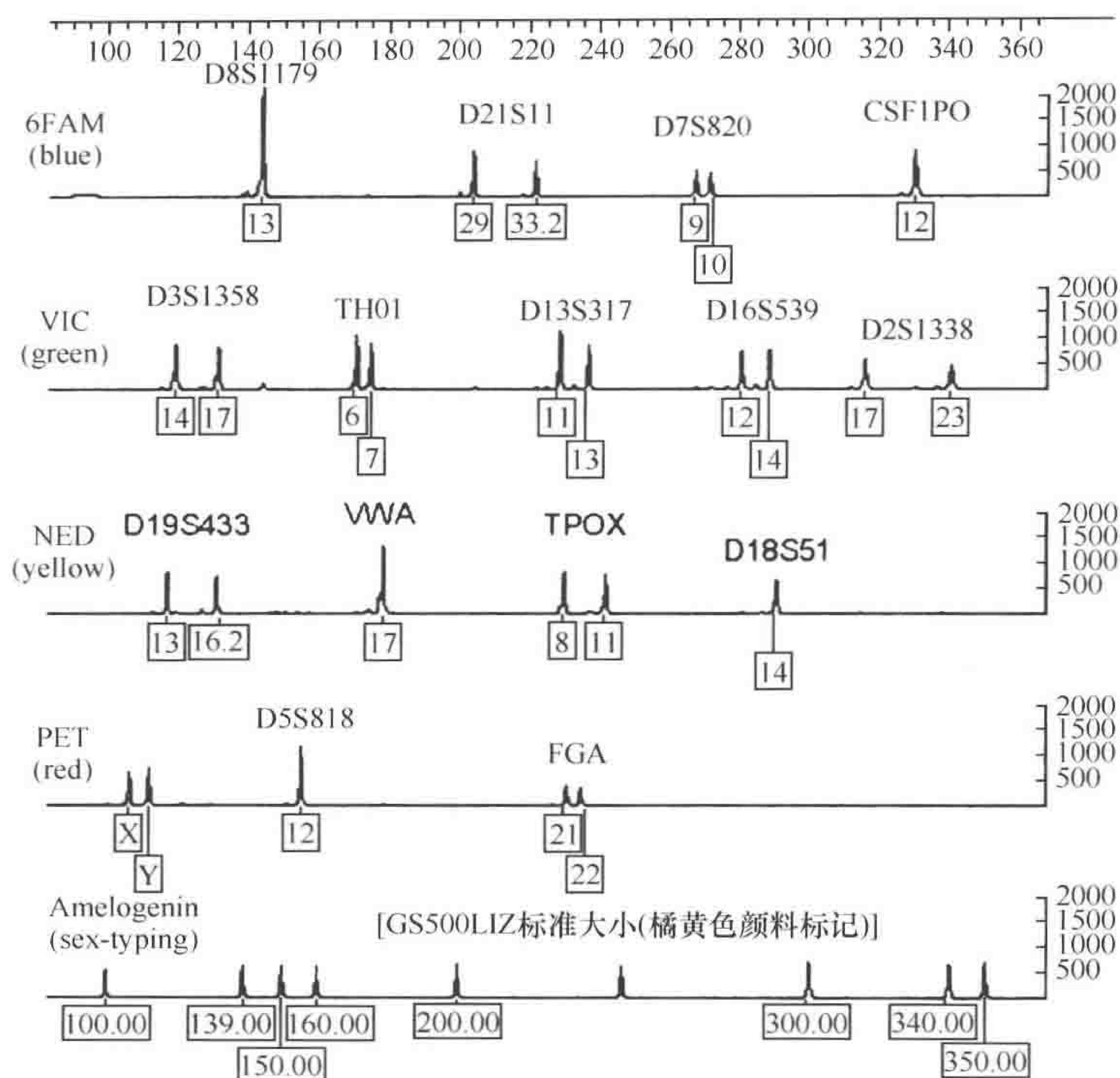



图 34-4 用 Identifier 试剂盒得到的 STR 分型的电泳图谱与第二组例子进行比对。列在下面的基因型对每个 STR 位点表示出一个峰（见图版）

一系列生物学上的和仪器上的人造物必须经常挑拣出来以产生完备、精确的 STR 结果（SWGDAM 2000；也可参见第 6 章和第 15 章，Butler 2005）。生物学的人造物包括模糊的产物、由腺苷酸化不完全造成的劈裂的峰、三等位基因类型和带有在重复区域或空白区域的突变的变异等位基因，该等位基因导致一个等位“离开了阶梯”。仪器上的人造物源于电压不稳、染色斑迹和染料颜色间的相互渗透等。样品的混合也可能使事情变得复杂，当两个或三个个体的不同成分在 STR 结果中同时存在时，完全的解码可能是困难的。

## 统计分析

一旦对特定样品得到了 DNA 结果，它就应放在整个案件中进行考虑。DNA 结果本身不会产生价值，它们必须与标准样品作比较。法庭案例的证据与来自嫌疑人或者（如果没有嫌疑人）过去已证实的罪犯的 DNA 资料库进行比对。从失踪者或社会灾难遇害者的遗留物中得到的 DNA 结果与生物学的亲属或直接的基准样品（如从牙刷上得到的 DNA）作比较。军事行动伤亡者的遗留物与个体入伍时留下的血斑作直接比较。在父权实验中从待选父亲得到的 DNA 结果与孩子的结果比较（在可能时也与孩子母亲的结果比较，为的是确定关键的等位确实来自父亲）。图 34-5 显示了从一个父亲、母亲

和三个孩子所做的亲子鉴定的结果。在孩子的基因型中可以看出STR等位的孟德尔遗传。例如，第一个孩子从她母亲那遗传了“12”等位基因，从父亲那遗传了“14”等位基因。

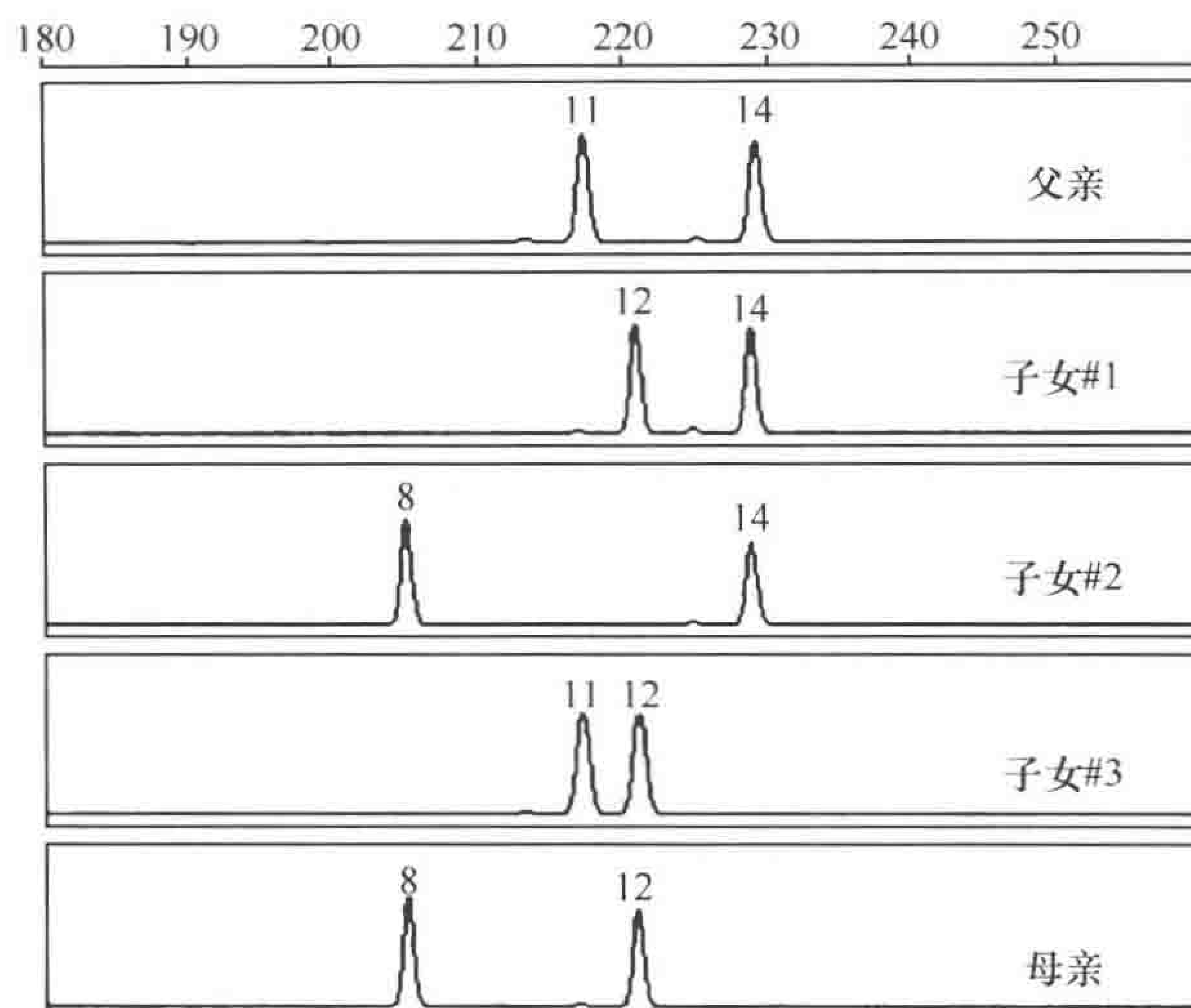



图 34-5　一个家庭的D13S317位点的STR构成显示了等位基因在3个孩子中的孟德尔式遗传

从DNA比对中可能产生三个结论：①匹配或包含；②不匹配或排除；③没有足够的资料可用使得结论不完全。如果观察到一个匹配，那么通常要给出该特定构成的稀有性估计。换句话说，所观察到的匹配有多大显著性？这种特定结构的稀有性估计，有时被称为随机匹配可能性（random match probability，RMP），是基于相关群体的等位基因频率。等位基因频率的信息是通过检测随机的，通常是100个或更多个体（他们自行确定其种族成分），并且对STR位点上所观察的每一种等位出现的次数进行计数而得到。然后对每一个位点的基因型频率依据哈代-温伯格平衡假设进行估计，辅以必要的修正以适合亚群体的可能结构。典型的情况是，法庭实验室利用由FBI实验室提供的称为PopStats的计算机程序来进行RMP计算，该程序伴有一个DNA索引系统（CODIS）软件。美国高加索人、非裔美国人和西班牙群体最经常被包含在基因频率引出的报告中，这些基因频率是根据由FBI组织的大群体研究得到的（Budowle et al. 2001)。需特别注意的是使用从不同群体研究得到的基因频率可能导致稍微不同的RMP。例如，上述第一组资料根据一个美国高加索群体研究（Butler et al. 2003a）得到的等位基因频率，得出的估计是$8.37\times10^{14}$分之一的RMP；但根据另外一个高加索资料库得出的估计是$2.46\times10^{15}$分之一的RMP（Budowle et al. 2001）。

## 时间考虑

上述简介的样品处理步骤通常需要1～2d，但也可能约5h就完成，主要的时间花在PCR热循环上。使用快速循环法STR可在1h之内分型（Belgrader et al. 1998）。

然而快速循环在大的多重扩增反应中效果不好。使用一个单一16毛细管ABI3100，一个操作者可以在1d内轻易地从100个以上的DNA样品中取得资料。许多法庭实验室正在引入机器人吸液管工作站（Robotic pipetting workstation）以便自动成批处理样品（图34-6和图34-7）。一些实验室使用带条形码的试管和平皿以及实验室信息管理系统（LIMS），使之能够做样品跟踪，有助于资料的有序保管。在整个操作中的一个主要的瓶颈是资料回顾。在许多实验室正在开始实现专家系统软件以加快资料回顾和STR整体解析的速度。法庭DNA测试资源的概要请参见表34-3。

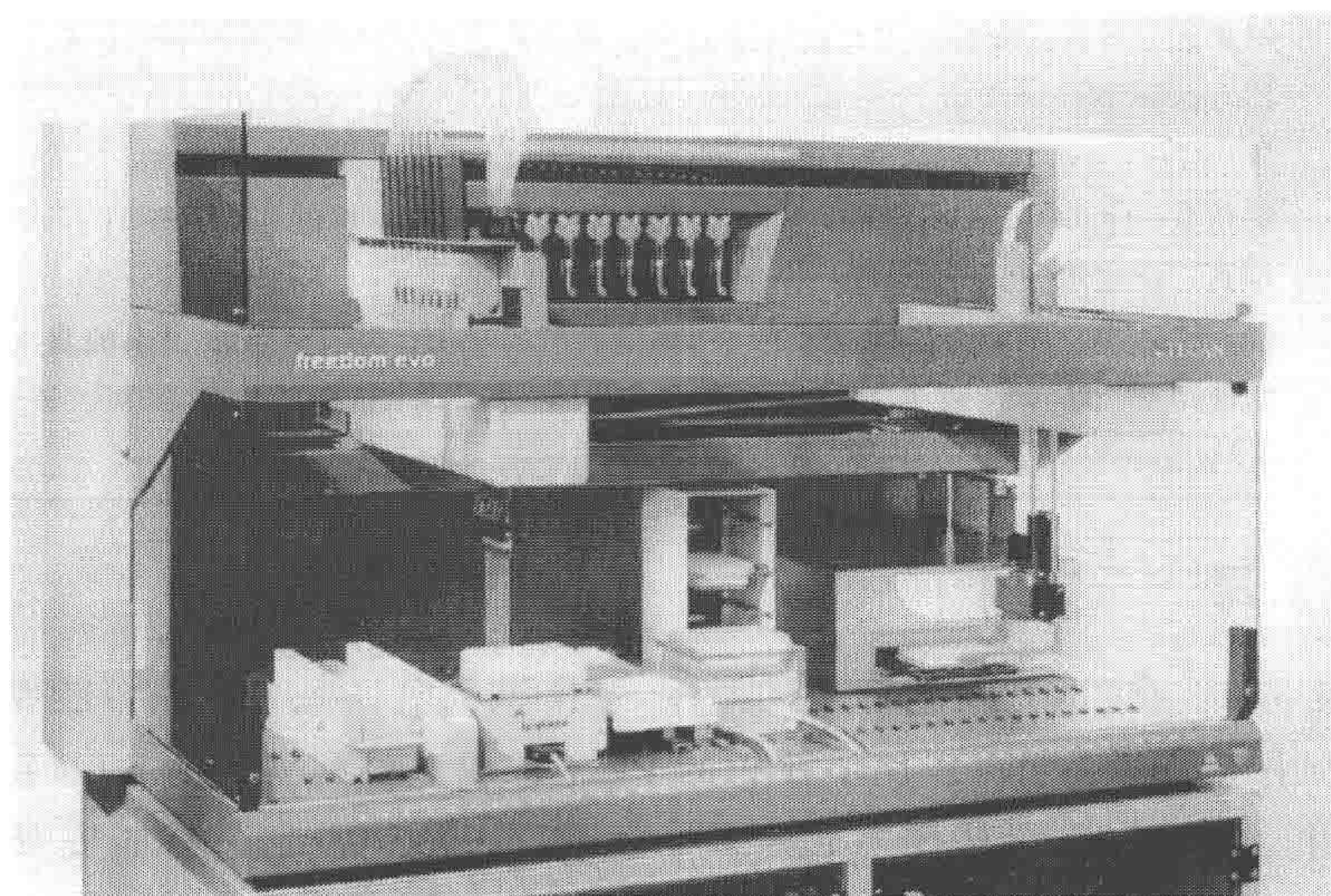



图34-6 执行STR分型手工步骤的机器人液态操作者。该图显示了可以增强实验室样品处理和分析能力的自动化EVO法庭工作站（得到Tecan Schweiz AG的允许）

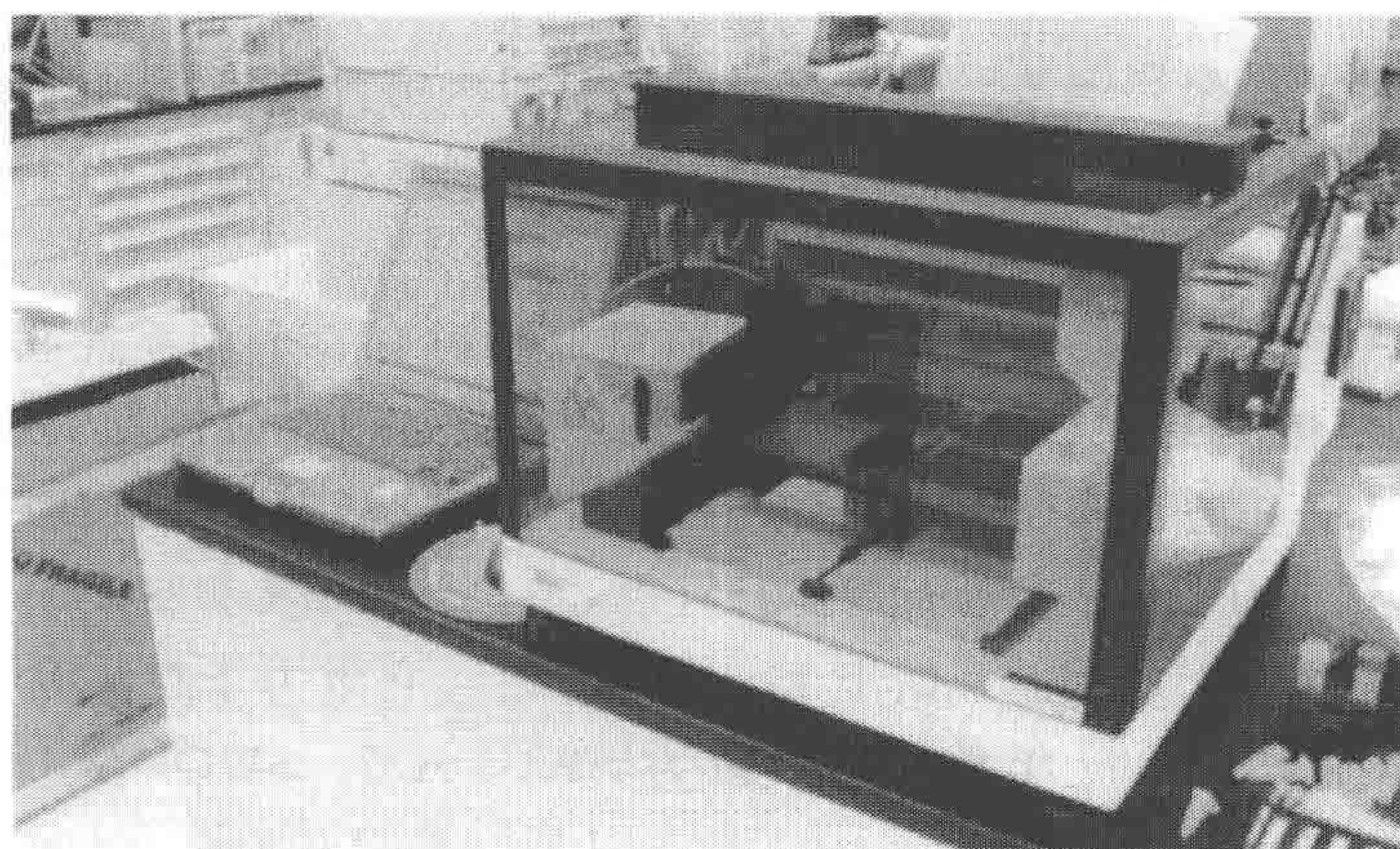


图34-7 在NIST实验室的Corbett机器人

表 34-3　DNA 测试的电子资源

| | |
|---|---|
| 数据库 | OmniPop 程序用于计算 RMP<br>Yfiler 数据库用于计算单倍型频率 |
| 软件 | FSS-i3 专家系统软件用于自动 STR 分型<br>GeneMapperID v3. 2<br>Sequencer 用于 mtDNA 序列分析 |
| 用于匹配概率计算的软件的网点 | Canadian Random Match Calculator（http://www. csfs. ca/pplus/profiler. htm）可以对 Applied Biosystems 出售的 Profiler Plus and COfiler 试剂盒所扩增的 13 个美国核心 STR 位点的 STR 结果进行计算。该程序可比较有限的 FBI 和加拿大人收集的基因频率的结果<br>European Network of Forensic Science Institutes DNA Working Group STR Population Database（http://www. str-base. org/index. php）使用从 24 个欧洲群体来的 5699 个样品进行 SGM Plus 试剂盒（Applied Biosystems）的 10 个 STR 位点的 STR 资料的匹配概率计算，该试剂盒在欧洲得到广泛应用<br>OmniPop 是由圣地亚哥警察部的一个法庭科学家 Brian Burritt 发展出来的基于 Excel 的程序。它使用从 202 个发表的数据库的等位基因频率计算 STR 结果的频率。该程序在 http://www. cstl. nist. gov/biotech/strbase/populationdata. htm 上免费使用 |

## STR 分型的费用

使用标准 50μL 或 25μL 反应体系的商用 STR 试剂盒，每个样品可能要花费 30～35 美元。有几个小组已经表明可以用小体积的 PCR 常规地得到可靠的 STR 结果，由此减少试剂的费用。皇家加拿大骑警（LeClair et al.　2003）和北路易斯安那犯罪实验室（Gaines et al.　2002）发表了使用商用 STR 试剂盒的 5μL PCR 方案。人力仍然是整个实验过程中最昂贵的部分。对于父权鉴定或法庭案件所做的 STR 分型，一个契约实验室对每个样品一般收取几百美元的费用。

## 法庭 DNA 测试的质量保障

法庭 DNA 测试结果影响法庭作出有罪或无罪的决定，这将影响到被告的自由。今天，以科学为基础的法庭 DNA 测试在法庭上极少受到挑战。实验室的操作和所进行的分析的典型评价表明了 DNA 测试的可靠性。

在说明 DNA 测试结果的可靠性和维持质量保障时有 5 个相互关联的部分。第一，质量保障标准已经由 FBI 主管（FBI　2000a，b）发行，以有效性和训练为主题的指导手册也已由 DNA 分析方法科学工作小组（SWGDAM）制定出来。这些标准和指导给出了基准并管理了法庭 DNA 实验室的运行（表 34-4）。FBI 实验室发起的 DNA 分析方法技术工作小组（TWGDAM）开始于 1988 年 11 月，它致力于在法庭 DNA 测试过程中的质量保障事宜。在 1998 年它的名字改为 SWGDAM。SWGDAM 每年开两次会，通常是 1 月和 7 月在弗吉尼亚的 Quantico 的 FBI 研究院召开。它由 40～50 名全美国的法庭实验室的 DNA 技术指导和科学家组成，他们在各附属委员会中对不同主题制定指导方针。然后这些指导方针在 *Forensic Science Communications* 上发表，可以在 FBI. gov 网站上找到（早期的 TWGDAM 资料发表在 *Crime Laboratory Digest* 上）。

DNA顾问委员会（DAB）是一个有13个投票成员的议会委托的实体，它存在于1995～2000年，负责为法庭DNA团体颁布标准。自从2000年，SWGDAM作为负责为美国国内法庭社团提供建议的小组开始运行。FBI实验室监督所有参加Combined DNA Index System（CODIS）的实验室审计。第二，对实验室予以授权（并且有规律地重新授权），这表明这些实验室已根据质量保障标准发展出标准的操作步骤（SOP），并且在日常工作中遵循这些步骤。犯罪实验室由美国犯罪实验室指导协会-实验室指派委员会（American Society of Crime Laboratory Directors-Laboratory Accreditation Board，ASCLD-LAB）指派。只管理DNA数据库，不进行个案操作的契约实验室可以由国家法庭科学技术中心（National Forensic Science Technology Center，NFSTC）任命。第三，每半年进行一次熟练性测试，以确保分析者有能力根据实验室SOP进行可靠的试验。第四，通过已被证明的参考材料，如国家标准和技术研究所（National Institute of Standards and Technology，NIST）的标准参考物（SRM），进行设备校准和操作方案的验证。第五，每年或每半年进行视察或实验室审查以确保质量保证标准被遵守，这些标准包括分析者的熟练测试和仪器/操作方法的校准。

**表34-4　SWGDAM建立的标准和指南**

| | |
|---|---|
| TWGDAM指南 | Kearney等.（1989）Guidelines for a quality assurance program for DNA restriction fragment length polymorphism analysis. *Crime Lab Digest* 16：40-59<br>Kearney等.（1991）Guidelines for a quality assurance program for DNA analysis. *Crime Lab Digest* 18：44-75.<br>Budowle等.（1995）Guidelines for a quality assurance program for DNA analysis. *Crime Lab Digest* 22：20-43. |
| DAB标准 | http://www.fbi.gov/hq/lab/fsc/backissu/july2000/codis2a.htm Quality Assurance Standards for Forensic DNA Testing Laboratories (issued by the DNA Advisory Board July 1998)<br>http://www.fbi.gov/hq/lab/fsc/backissu/july2000/codis1a.htm Quality Assurance Standards for Convicted Offender DNA Databasing Laboratories (issued by the DNA Advisory Board April 1999)<br>http://www.fbi.gov/hq/lab/fsc/backissu/july2000/dnastat.htm. Statistical and Population Genetics Issues Affecting the Evaluation of the Frequency of Occurrence of DNA Profiles Calculated from Pertinent Population Database(s) (issued by the DNA Advisory Board February 2000) |
| SWGDAM指南 | http://www.fbi.gov/hq/lab/fsc/backissu/april2003/swgdambylaws.htm Bylaws of the Scientific Working Group on DNA Analysis Methods<br>http://www.fbi.gov/hq/lab/fsc/backissu/april2003/swgdammitodna.htm Guidelines for Mitochondrial DNA (mtDNA) Nucleotide Sequence Interpretation<br>http://www.fbi.gov/hq/lab/fsc/backissu/april2003/swgdamsafety.htm Guidance Document for Implementing Health and Safety Programs in DNA Laboratories<br>http://www.fbi.gov/hq/lab/fsc/backissu/oct2001//kzinski.htm Training Guidelines<br>http://www.fbi.gov/hq/lab/fsc/backissu/july2000/strig.htm Short Tandem Repeat (STR) Interpretation Guidelines<br>http://www.fbi.gov/hq/lab/fsc/backissu/july2004/standards/2004_03_standards02.htm Revised Validation Guidelines<br>http://www.fbi.gov/hq/lab/fsc/backissu/july2004/standards/2004_03_standards03.htm Report on the Current Activities of the Scientific Working Group on DNA Analysis Methods Y-STR Subcommittee (recommended core Y-STR loci) |
| 所有的法庭DNA实验室的审核文件 | http://www.fbi.gov/filelink.html? file =/hq/lab/fsc/backissu/july2004/pdfs/seubert.pdf Quality Assurance Audit for Forensic DNA and Convicted Offender DNA Databasing Laboratories (issued by the FBI Laboratory July 2004) |

# 数据问题

## 生物学假象

在评估 STR 数据时，常见生物学假象包括：模糊的产物、腺苷酸化不完全、变异的等位和三等位基因类型（Butler 2005，第 6 章）。模糊的产物源于 PCR 扩增时的重复片段上链的滑动。对于 4 核苷酸重复位点，模糊的产物可以是比正常等位短 4 个碱基，通常小于 10%的等位基因高度。腺苷酸化不完全导致劈裂的峰，每个等位有一个“－A”和一个“＋A”产物。“－A”峰是全长的 PCR 产物，“＋A”峰则源于 DNA 聚合酶产生的多余碱基。不是所有聚合酶都会增加这个多余的核苷酸，但这不能说明什么，它的产生频率依赖于 PCR 产物的 3′端，该末端由反向引物的 5′端所限定（Butler 2005，表 6-1）。变异的等位是由于在重复区域内部或附近插入或缺失一个核苷酸所产生的。三等位基因类型最常在 TPOX 和 D18S51 中观察到，可能源于被扩增两条染色体之一的引物结合位点的重复。NIST 维护着一个用于人类鉴定实验的 STR 标记的互联网数据库，称为 STRBase（http://www.cstl.nist.gov/biotech/strbase）。STRBase 将变异等位基因和在世界上的法庭 DNA 科学家所观察到的三等位基因类型编入目录。

## 降解的 DNA 材料

当 DNA 已被降解成小片段，就不能有效地形成大的 PCR 产物，这是因为模板区域有可能已是碎片。由于 DNA 样品损害时 STR 位点周围完整的、未损伤的分子减少，PCR 产物变长时会提示有信号损失。使引物靠近重复区域以产生小的 STR 系统可以改善用降解的 DNA 样品进行的扩增效果（Butler et al. 2003a）。

## 混合

在法庭案例中有时样品中混合了两个或多个个体的材料，特别是在性袭击证据中。几乎等量的 DNA 的混合可以通过多位点上两个以上的等位的存在而较容易被察觉，如图 34-8 所示。对所有提供者全部 STR 资料的解码是一项挑战性的工作，通常要依靠已

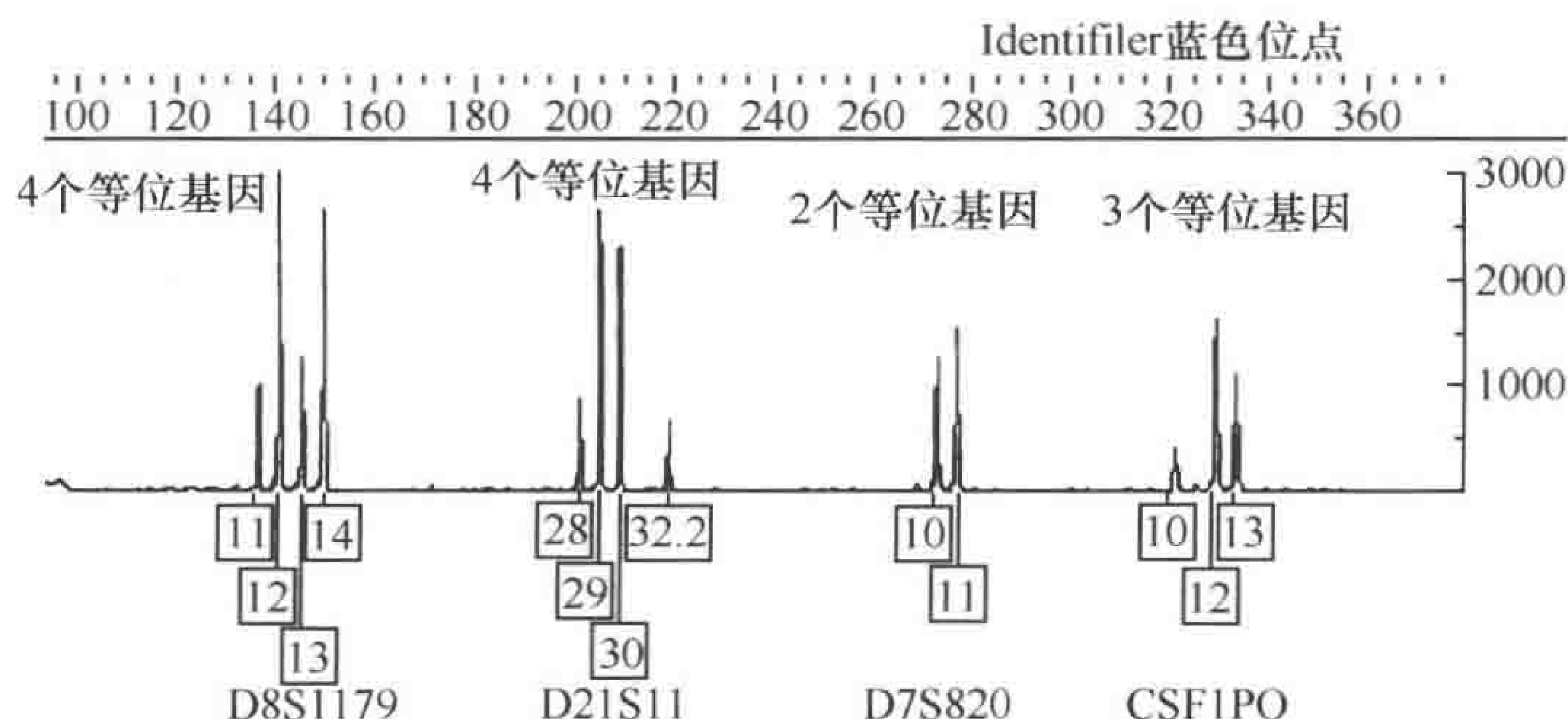



图 34-8　对一个混合 DNA 样品进行分析的 4 个 Identifiler 蓝色标记的 STR 位点的地区图。注意在 4 个所显示的位点中有三个具有两个以上的等位基因，提示在 DNA 材料中含有不止一个个体的样品（见图版）

知材料的帮助，如受害者的资料。由于 4 或 5 核苷酸标记产生模糊的产物的可能性较小，而模糊的产物造成对混合的 DNA 材料的解读困难，因此法庭的测试最好采用 4 或 5 核苷酸标记（Butler 2005，第 7 章）。

## 其他的法庭 DNA 测试技术

### Y 染色体 STR 测试

Y 染色体只存在于男性个体，因此 Y-STR 测试可以从混合了两性 DNA 的样品中发现男性特异性信息，甚至当女性 DNA 量超过男性上千倍时也可做到这一点（Butler 2005，第 9 章）。在 Y 染色体上已经鉴定了几百个 STR，但只有十几个用于人类鉴定实验（Butler 2003）。这些 Y-STR 都位于 Y 染色体的非重组区，这使它们在区分两个不相干个体时比一套常染色体标记差些，并且不能将父子或兄弟彼此区分开（除非产生了突变）。

### 线粒体 DNA

在测试高度降解的样品，或样品含有很少的 DNA，如稳定期的毛干时，法庭科学家常转而求助于线粒体 DNA（mtDNA）分析（Butler 2005，第 10 章）。细胞中的 mtDNA 比核 DNA 的拷贝数高。在细胞中每个 STR 位点只有两个拷贝，但 mtDNA 有成百个拷贝。人类的线粒体只遗传自个体的母亲，因此它不是独特的。只要没有突变发生，同胞、姨妈和其他母系亲属都具有同样的 mtDNA 序列。通常测序在 mtDNA 控制区内的两个高变区，与被称为修订的剑桥参考序列（图 34-9）的标准序列相比，检测到了 610bp（mtDNA 基因组通常存在 16568bp）的差异（Anderson et al. 198l；Andrews et al. 1999）。mtDNA 序列信息分析费时费力，也比 STR 分型昂贵，但当用传统的 STR 分型方法无效时，它通常能产生结果。

### SNP 用于估计种族和表型特征

SNP 将很可能是补充而不是替代目前的 STR 分型系统（Gill et al. 2004）。因为 SNP 较 STR 突变率低（$10^{-8}$ vs $10^{-3}$）。在遗传变异的特定群体中 SNP 更可能被当作是“固定的”，因此可根据少量的遗传差异进行种族估计。目前正在运用 SNP 区分发色、眼睛颜色及其他的表型性状（Butler 2005，第 8 章）。多数这些实验仍处于研究的不同阶段，远未达到常规的人类鉴定实验阶段。

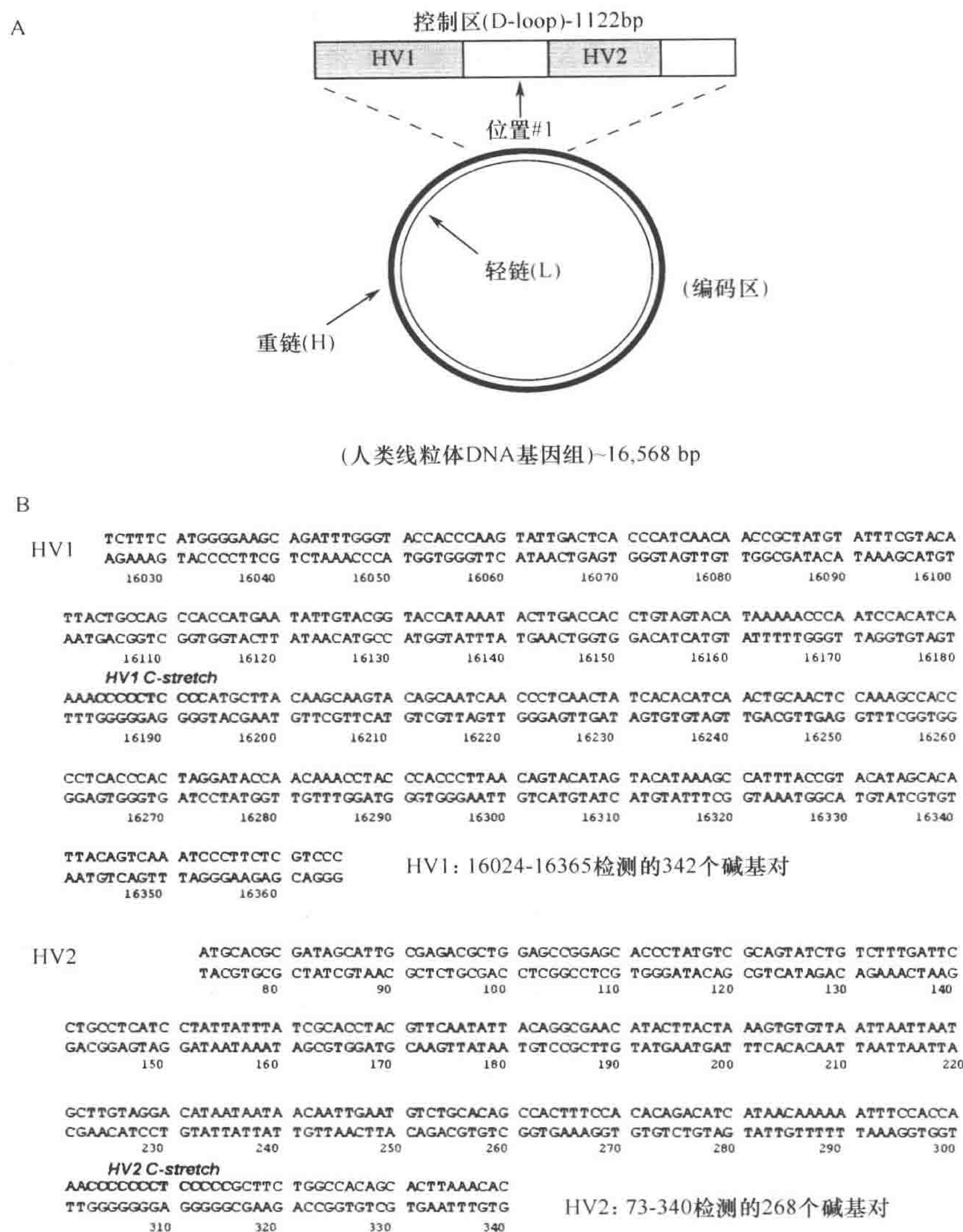



图 34-9　A. 表示人类 mtDNA 基因组示意图，显示编码区和控制区的相对位置，以及两个通常在法庭 DNA 测试时检验的高变（HV）部分。B. 表示存在于人类 mtDNA HVl 和 HV2 的修订的剑桥参考序列的 610bp 的标准序列。样品结果典型地用与参考样品的差别（位置和核苷酸的改变）来表示

## 概要

STR 已经并且在可见的将来仍将是在人类鉴定实验中检测遗传变异的主要方法（NIJ　2000）。一套不超过 24 种的核心 STR 位点在法庭和父权鉴定实验方案中占主导地位（Butler　2006）。每年使用这些核心 STR 位点产生了数以百万计的 DNA 鉴定结果，服务于人类鉴定的目的。

# 致谢

本工作通过跨机构协议（Interagency Agreement）2003-IJ-R-029 和法律实施标准 NIST 办公室得到了国家司法研究所（National Institute of Justice）的部分支持。非常感谢 NIST 人类鉴定项目组成员 Amy Decker、Dave Duewer、Becky Hill、Margaret Kline、Jan Redman 和 Peter Vallone 的协助。所写均为作者的观点，不代表美国司法部。指明特定的商业设备、仪器和材料，是为了尽可能详细地说明实验步骤。这些说明并不意味着推荐或为国家标准和技术研究所所认同，也不意味着所有这些材料、仪器或设备肯定最适合实验目的。

# 参考文献

AABB (American Association of Blood Banks). 2005. Annual report summary for testing in 2004 prepared by the Relationship Testing Program Unit. Available at http://www.aabb.org/Documents/Accreditation/Parentage_Testing_Accreditation_Program/rtannrpt04.pdf

Anderson S., Bankier A.T., Barrell B.G., de Bruijn M.H., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., et al. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290:** 457–465.

Andrews R.M., Kubacka I., Chinnery P.F., Lightowlers R.N., Turnbull D.M., and Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genet.* **23:** 147.

Belgrader P., Smith J.K., Weedn V.W., and Northrup M.A. 1998. Rapid PCR for identity testing using a battery-powered miniature thermal cycler. *J. Forensic Sci.* **43:** 315–319.

Biesecker L.G., Bailey-Wilson J.E., Ballantyne J., Baum H., Bieber F.R., Brenner C., Budowle B., Butler J.M., Carmody G., Conneally P.M., et al. 2005. DNA identifications after the 9/11 World Trade Center attack. *Science* **310:** 1122–1123.

Brown K. 2002. Tangled roots? Genetics meets genealogy. *Science* **295:** 1634–1635.

Budowle B., Smith J., Moretti T., and DiZinno J. 2000. *DNA typing protocols: Molecular biology and forensic analysis.* Eaton Publishing, Natick, Massachusetts.

Budowle B., Shea B., Niezgoda S., and Chakraborty R. 2001. CODIS STR loci data from 41 sample populations. *J. Forensic Sci.* **46:** 453–489.

Butler J.M. 2003. Recent developments in Y-short tandem repeat and Y-single nucleotide polymorphism analysis. *Forensic Sci. Rev.* **15:** 91–111.

———. 2005. *Forensic DNA typing: Biology, technology, and genetics of STR markers.* Elsevier, New York.

———. 2006. Genetics and genomics of core STR loci used in human identity testing. *J. Forensic Sci.* **51:** 253–265.

Butler J.M., Shen Y., and McCord B.R. 2003b. The development of reduced size STR amplicons as tools for analysis of degraded DNA. *J. Forensic Sci.* **48:** 1054–1064.

Butler J.M., Buel E., Crivellente F., and McCord B.R. 2004. Forensic DNA typing by capillary electrophoresis: Using the ABI Prism 310 and 3100 Genetic Analyzers for STR analysis. *Electrophoresis* **25:** 1397–1412.

Butler J.M., Schoske R., Vallone P.M., Redman J.W., and Kline M.C. 2003a. Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. *J. Forensic Sci.* **48:** 908–911.

Collins P.J., Hennessy L.K., Leibelt C.S., Roby R.K., Reeder D.J., and Foxall P.A. 2004. Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: The AmpFlSTR Identifiler PCR Amplification Kit. *J. Forensic Sci.* **49:** 1265–1277.

Comey C.T., Koons B.W., Presley K.W., Smerick J.B., Sobieralski C.A., Stanley D.M., and Baechtel F.S. 1994. DNA extraction strategies for amplified fragment length polymorphism analysis. *J. Forensic Sci.* **39:** 1254–1269.

FBI (Federal Bureau of Investigation). 2000a. Quality assurance standards for forensic DNA testing laboratories. *Forensic Sci. Commun.* **2:** no. 3 (July 2000A). Available online: http://www.fbi.gov/hq/lab/fsc/backissu/july2000/codispre. htm.

FBI (Federal Bureau of Investigation). 2000b. Quality assurance standards for convicted offender DNA databasing laboratories. *Forensic Sci. Commun.* **2:** no. 3 (July 2000B). Available online: http://www.fbi.gov/hq/lab/fsc/backissu/july2000/codispre.htm.

Foster E.A., Jobling M.A., Taylor P.G., Donnelly P., de Knijff P., Mieremet R., Zerjal T., and Tyler-Smith C. 1998. Jefferson fathered slave's last child. *Nature* **396:** 27–28.

Gaines M.L., Wojtkiewicz P.W., Valentine J.A., and Brown C.L. 2002. Reduced volume PCR amplification reactions using the AmpFlSTR Profiler Plus kit. *J. Forensic Sci.* **47:** 1224–1237.

Gill P. 2002. Role of short tandem repeat DNA in forensic casework in the UK—Past, present, and future perspectives. *BioTechniques* **32:** 366–372.

Gill P., Jeffreys A.J., and Werrett D.J. 1985. Forensic application of DNA 'fingerprints'. *Nature* **318:** 577–579.

Gill P., Werrett D.J., Budowle B., and Guerrieri R. 2004. An assessment of whether SNPs will replace STRs in national DNA databases: Joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM). *Sci. Justice* **44:** 51–53.

Gill P., Ivanov P.L., Kimpton C., Piercy R., Benson N., Tully G., Evett I., Hagelberg E., and Sullivan K. 1994. Identification of the remains of the Romanov family by DNA analysis. *Nat. Genet.* **6:** 130–135.

Jeffreys A.J., Wilson V., and Thein S.L. 1985. Hypervariable 'minisatellite' regions in human DNA. *Nature* **314:** 67–73.

Krenke B.E., Tereba A., Anderson S.J., Buel E., Culhane S., Finis C.J., Tomsey C.S., Zachetti J.M., Masibay A., Rabbach D.R., et al. 2002. Validation of a 16-locus fluorescent multiplex system. *J. Forensic Sci.* **47:** 773–785.

Leclair B., Sgueglia J.B., Wojtowicz P.C., Juston A.C., Fregeau C.J., and Fourney R.M. 2003. STR DNA typing: Increased sensitivity and efficient sample consumption using reduced PCR reaction volumes. *J. Forensic Sci.* **48:** 1001–1013.

NIJ (National Institute of Justice). 2000. The future of forensic DNA testing: Predictions of the Research and Development

Working Group of the National Commission on the Future of DNA Evidence. Washington, D.C. http://www.ojp.usdoj.gov/nij/pubs-sum/183697.htm

———.2006. Lessons learned from 9/11: DNA identification in mass fatality incidents. Washington, D.C. NCJ214781. Available at http://massfatality.dna.gov

SWGDAM (Scientific Working Group on DNA Analysis Methods). 2000. Short tandem repeat (STR) interpretation guidelines. *Forensic Sci. Commun.* 2(3): on-line at http://www.fbi.gov/hq/lab/fsc/backissu/july2000/strig.htm

Vanek D., Hradil R., and Budowle B. 2001. Czech population data on 10 short tandem repeat loci of SGM Plus STR system kit using DNA purified in FTA cards. *Forensic Sci. Int.* **119:** 107–108.

Walsh P.S., Metzger D.A., and Higuchi R. 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *BioTechniques* **10:** 506–513.

Whitaker J.P., Clayton T.M., Urquhart A.J., Millican E.S., Downes T.J., Kimpton C.P., and Gill P. 1995. Short tandem repeat typing of bodies from a mass disaster: High success rate and characteristic amplification patterns in highly degraded samples. *BioTechniques* **18:** 670–677.

# 35 人类基因组：前面是什么

Michael P. Weiner[1] and J. Claiborne Stephens[2]

[1]RainDance Technologies, Guilford, Connecticut 06437; [2]Motif BioSciences Inc., New York, New York 10017

**技术的进步**
**测序的应用和意义**
**生物多样性的变化**

## 技术的进步

许多分子生物学技术进展快速，我们10年前使用的技术与今天使用的大不相同。在许多情况下这种变化代表了一种改善，它使目前的方法更快、更便宜、更加自动化。然而，技术偶然地呈现一种跃进式的变化，类似于一种点断平衡。DNA测序技术的历史以及主要的改善可以说明这一点。1977年，在英国医学研究委员会的Fred Sanger发明了一种运用链终止的DNA测序方法。由DNA聚合酶、放射性标记、链终止的双脱氧核苷酸和聚丙烯酰胺凝胶电泳所构成的方法在世界范围内被广泛地接受。约10年之后（1986年）Leroy Hood和他的同事发明了一种机器自动进行部分的Sanger测序步骤，并且用荧光染料替代了放射性物质。第一个基于CaI Tech设计的DNA测序仪器出现在1年以后（1987年），并获得了商业上的成功。随着这些仪器的采用，研究者可以对小的细菌基因组和简单的生物（酵母、果蝇和蠕虫）基因组进行测序。但测序人类基因组所需的劳务费用仍然太高，拼凑整个基因组还需要等到不同的技术出现才能进行。1998年5月，随着基于第一个96道毛细管电泳的DNA测序仪的引入，这个技术终于在11年后到来。该技术的使用以及它所伴随的费用的减少使我们能够在3年后（2001年）完成人类基因组测序草图。到2007年下半年，新一代的测序仪已经又使用了几年。于2004年由454 Life Sciences和2007年由Solexa引入的这类仪器使用大规模平行合成测序方法，一次可解读上亿乃至千亿的碱基序列。

测序技术进步与这个人工操作的基因组的变异课题有什么关系？答案是测序技术和基因组变异技术正开始融合，这一点正变得越来越明显。全基因组测序的费用正与全基因组SNP分析的费用接近。目前的技术可能不会把我们带到那里，但我们知道下一代技术将会做到，这一点令人兴奋。

## 测序的应用和意义

经过多年的合作努力，人类基因组计划第一次揭示了人类基因组序列，这是一个不

同基因组的嵌套。正像我们刚讨论过的，超高生产量的测序平台可在一次操作中解读上亿的碱基，因此极大地减少了时间消耗。还值得注意的是，这些进步极大地降低了错误率（典型地少于 1/20 000），提供了高度的个体阅读的精确性。为这些有希望的结果所鼓励，有两个公司已经开始了测序个体基因组的项目。454 Life Sciences 已经宣布为一个 78 岁的白种男人的 DNA 进行初步测序，他就是詹姆斯-沃森。选择沃森是一种好奇心的满足——当然，他因为在发现 DNA 结构中所扮的角色为人所知，沃森-克里克 DNA 双螺旋的沃森，他也是人类基因组计划的第一任主席。在另一个相关的计划中，Illumina 正着手对约鲁巴，一个尼日利亚部族的个体进行测序，这是由于 HapMap 计划的选择。在 HapMap 计划中，对 4 个群体的遗传多样性进行研究：尼日利亚伊巴丹的约鲁巴人、北京的汉族中国人、东京的日本人和法国的 Centre d'Etude du Polymorphisme Humaine（CEPH）资源。HapMap 计划正在继续，并且延伸了人类基因组计划的精神。这些单倍型的研究提示约鲁巴人在人类遗传多样性方面比其他群体有较大程度或更完全的表达——显然，对这个群体进行基因组测序将比参考群体更有意义。这些变异研究标本的取得遵循了适当的道德委员会所批准的程序，并依据社团约定强调特殊的关切和获得广泛的同意，这一点很重要（对这些步骤和要求的更广泛的讨论见第 1 章）。然而这些仍存在伦理问题。通过样品与个体鉴定者脱钩使隐私得到保护，但是与种族群体存在的潜在关联或暗示受到关注，这些我们在下面详细讨论。

这些测序和基因组变异的研究代表了第一次探究个体基因组与他或她生物学性状和倾向之间关系的尝试。我们期待近期测序技术的改善可以使基因组分析像血液检查一样成为常规。这具有明确和潜在的利益，特别是在医学前沿，如筛查与不同类型癌症相关的突变。原则上，这些基因组信息可以帮助内科医生进行个体化的（因而是最优化的）治疗，以确保更好的健康或医疗效果。然而还存在对误解，或更严重地，误用这些信息的担忧。我们得到基因组资料的能力对医学下游的所有方面施加了巨大的压力，使之跟上萌生的资料量。每个在卫生保健网中的赌金保管者——患者、医师、付款人、保险公司、调整者、教育者等。希望建立一个法律上有效的遗传测试，否则目前的下游结构不可能有能力跟上可用资料的步伐。社会，特别是需要批准和调节遗传测试的那些机构，如何才能跟上可能的实用性和需求呢？

## 生物多样性的变化

我们已经讨论了在人类群体中与遗传变异有关的事项，并考虑了一些可能的前景。然而，如果延伸我们的讨论至这个星球的其他物种的话，显然我们正在目睹一些发生在植物和动物王国的快速的（在一些例子中，令人懊丧的）变异性改变。

在有大量追随者的科幻电影 Silent Running 中，在没有森林的地球上，生物多样性的存留物只是一些离开世界的带有自我维持生物圈的飞船。这些飞船的使命被认为是没必要的，它自己的生物圈也被抛弃在太空。该电影在 20 世纪 70 年代早期发行，那时生态学和环境都处于公众关注的焦点。亚马逊盆地森林的毁坏和影响到珊瑚礁存活的污染的增加受到了广泛的公众关注。

30 年后，在写这本书的时候，地球生物多样性的减少再次成为公众关注的焦点，对保留我们的生物多样性的关心已经开始超越国界。作为对这种关心的回应，挪威采取了一个有远见的举措——在北极建筑了一个“世界末日库”，储备植物种子以防备地球上可能的大灾难。一系列的研究越来越清楚地表明，是人类的活动导致了这个星球温度的升高，这种升高使得生物多样性减少。但是全球温度的升高仅仅是造成生物多样性降低的原因之一。我们日益广泛采用的农业技术无意识造成的结果是均一化的高产粮食和家畜，导致遗传多样性的丢失。

这一指南的几个章节描述了动物和植物物种正处于危险中，灭绝了，或接近灭绝。在历史上我们处于一个独特的位置，在这里我们正在学习生物体的遗传结构的种种细节，这些生物体正是由于我们的原因而处于危险和灭绝中。当一个物种消失时谁会知道我们在理解自己的方面损失了什么？当我们开始知道得更多，并开始赏识我们周围世界的物种遗传多样性后。我们就必须更努力地奋斗以保证它们的存活。

# 附录　注意事项

## 一般注意事项

请注意，本指南所附的注意事项并非详尽无遗。读者需要通过不断咨询各生产厂商以及利用其他资源来获取最新的和特殊的产品信息。对于在本书中涉及的有些化学物质和其他药品，我们并没有在实验方案中使用惊叹号（!）来标志其具有危害性。但是，如果操作不当的话，它们有可能给使用者带来伤害。请联系当地的安全机构或参考制造厂商的安全指导来获得更多信息。

无论何时都应该遵守下列注意事项。

- 在开始操作之前，要对即将**使用的所有物质的性质十分熟悉**。
- **没有提醒注意的物质**未必意味着是安全的，因为没有信息是完美无缺的。
- **如果接触了有毒物质**，要立即与当地安全机构联系，以接受指导。
- 对使用过的所有化学、生物和放射性废物要使用**适当的处理程序**。
- **对于合适手套的特殊准则**，向当地安全部门查询。
- **处理浓酸和浓碱要特别小心**。如果处理的量大时要戴护目镜、手套和合适的面罩。

强酸不能与有机溶剂混合，因为它们可能起化学反应。特别是硫酸和硝酸能发生剧烈放热反应，易引起大火和爆炸。

强碱不能与卤化物溶剂混合，因为它们能形成可导致爆炸的活性碳烯。

当由浓酸储存液制备酸的稀释液时要将酸注入水中。

- **小心操作和存放高压气体容器**，因为它们可能包含有易燃、有毒、腐蚀性气体以及窒息剂和氧化剂。并保证向销售商索要物质安全信息表，来获得正确的操作步骤。
- **移液管绝不能用嘴来吸取**。这种方法既不能保证无菌，也是危险的。一定要用移液器或吸液球。
- **将卤化物溶剂与非卤化物溶剂分开保存**（如氯仿和丙酮在碱性条件下混合会引起意外反应）。卤化物溶剂是有机溶剂，如氯仿、二氯甲烷、甲醇、苯、甲苯、N，N-二甲基甲酰胺（DMF）、二甲基亚砜（DMSO）和氰化甲烷。非卤化溶剂有戊烷、庚烷、乙醇、甲醇、苯、甲苯、DMF、DMSO 和乙腈。
- **激光辐射**，可见或不可见的激光都能引起对眼睛和皮肤的严重损伤。采取适当预防措施以防止接触直射和反射的光束。始终遵守制造商的安全指导准则并向当地安全机构咨询。更详细的信息参见下面的注意事项。
- **闪光灯**：由于其强烈的光对眼睛有害，有时也会发生爆炸。要戴适当的护目镜并遵守制造商的指导准则。

• **固定液和显影液和感光材料**均含有害化学物质。按照厂商的指导说明小心操作。

• **电源和电泳设备**：如果使用不当会引起严重火灾和电击事故。

• **实验室中的微波炉和高压灭菌锅**需要可靠的预防措施，它们在使用中（如熔化琼脂或带有琼脂的培养基，包括灭菌）可能会发生事故。如果瓶子上的螺旋口瓶盖没有拧松或没有松到足够的程度，蒸汽则没有足够空隙排出，这样当瓶子从微波炉或高压锅中拿出来时会引起爆炸。因此，在容器放入微波炉或高压锅前，务必将瓶盖拧松。另外，针对常规琼脂糖凝胶并不需要无菌操作，可以在称取琼脂后置于普通烧瓶中加热熔解。

**超声仪器的使用**。在使用高频率声波（16～100 kHz）用于细胞破碎和其他用途时。超声通过空气传播不会直接对人体产生伤害，但是结合大量的可听得见的声波会产生不同的反应，包括头痛、恶心、耳鸣等。因此，应避免身体直接接触高密度的超声波（不包括医疗影像设备）。使用时可应用适当的耳朵保护，并在使用超声设备的实验室房门上粘贴明显的警示标志。

• **使用刀片、解剖刀、剃刀或注射针**等剪切机械时要非常小心。刀片非常锋利！如果操作者使用不熟练，应找经验丰富的人演示正确的使用程序。为了处置得当，实验室里要配置一个专门的“利器”处理容器。丢弃用过的无防护罩的注射针时，应与连接的注射器一起丢掉。这种方法可以防止在使用用过的注射针头时发生的伤害（和可能的感染，见生物安全）。因为在试图把针头装回注射器时，往往会发生事故。注意：破损的巴斯德移液管、盖玻片或载玻片都可能造成伤害。

## 化学药品的一般特性

危险物质可按种类归结为下述目录。

• **无机酸（强酸）** 盐酸、硫酸、硝酸或磷酸这样的无机酸，大多是带有刺激烟雾的无色液体。避免溢出物洒到皮肤和衣服上。一旦溢出需要用大量清水及时冲洗。这些种类的浓酸能毁坏纸张、纺织品和皮肤，并能对眼睛造成严重伤害。

• **无机碱（强碱）** 例如，氢氧化钠等无机碱，呈白色固体形态，可溶于水，具有热挥发性。高浓度的碱溶液能缓慢溶解人的皮肤乃至指甲。

• **重金属盐** 通常是有色的固体粉末，可溶于水。许多重金属是强有力的酶抑制剂。因此，对人和环境（如鱼和水藻）是有毒性的。

• **有机溶剂** 大部分有机溶剂是易燃易挥发的液体。避免吸入有机溶剂的挥发气体，因其可引起恶心或者头晕。也要避免皮肤的接触。

• **其他的有机化合物** 包括像巯基乙醇之类的有机硫化合物或有机氨都有很重的令人讨厌的刺鼻气味。其他的高活性有机化合物也必须谨慎操作。

• **染料和染色液** 如果操作不当，染料及溶液不仅能使样品着色，也能污染皮肤和衣服。注意：某些染料（如溴化乙锭）是诱变剂、致癌剂，非常有害。

• **酶类** 几乎所有的以“ase”结尾的名词［如过氧化氢酶（catalase），β-葡萄糖醛酸糖苷酶（β-glucuronidase）或消解酶（zymolase）］都属于酶类。也有一些非系统命名的酶类，如胃蛋白酶（pepsin），是由制造厂商命名的。其制品中含有缓冲物质等，要

了解在这些物质中所包含原料的特性。

• **有毒化合物** 常用于处理细胞的化合物有一些是有毒化的（如环己酰亚胺、放线菌素 D 和利福霉素）。它们是有潜在危险的，要谨慎操作。

• 必须意识到以上所列化学物质的毒理学特性并未得到全面的研究。要以最合适的方法谨慎使用化合物。尽管化合物的毒性作用可以定量测定，如 $LD_{50}$ 值，但却不能判定其致癌或致突变的作用，特别是单一实验效应是不准确的。还应注意：危险性与化学物质的物理状态有关。（如微细粉末与晶体、二乙醚与甘油、干冰与气体罐中加压的二氧化碳之间的区别）。预期判断实验环境中合成产物最可能发生的结果，以及如何最大程度地保护你自己和周围的环境。

## 危险的物质

注意：原则上讲，专利药并未在此列出。大多数麻醉剂、染料、固定剂，以及试剂盒和其他商业化产品也未包含在内。麻醉剂需要特别注意。严格按照生产厂商的安全指南进行操作。

• **乙酸（浓）** 必须非常小心地操作。可能会由于呼吸、吞咽摄取或者皮肤吸收而受到伤害。因此，要戴合适的手套和护目镜，在化学通风橱里使用。

• **丙酮** 对皮肤、眼睛、黏膜和上呼吸道有刺激作用，勿吸入其气体。丙酮极其易燃，操作时佩戴合适的手套和护目镜，并远离热源、火花和明火。

• **氯化铵（$NH_4Cl$）** 可因吸入、咽下或皮肤吸收而危害健康。操作时要戴合适的手套和护目镜，在通风橱里进行。

• **硫酸铵［$(NH_4)_2SO_4$］** 可因吸入、咽下或皮肤吸收而危害健康。操作时佩戴合适的手套和护目镜。

• **溴化十六烷基三甲胺（CTAB）** 有毒，是一种兴奋剂，可因吸入、咽下或皮肤吸收而受到危害。操作时要戴合适的手套和护目镜。避免吸入其粉末。

• **氯仿（$CHCl_3$）** 对皮肤、眼睛、黏膜和呼吸道有刺激作用。它也是一种致癌剂，可损害肝和肾。避免吸入其挥发的气体。操作时戴合适的手套和护目镜，并始终在化学通风橱里进行。

• **二乙基焦磷酸胺（DEPC）** 是一种强有力的蛋白质变性剂，而且是可疑的致癌剂。开瓶时将瓶子远离身体，瓶内压力可能导致泼溅。操作时戴合适的手套，穿工作服，在化学通风橱里进行。

• **二硫苏糖醇（DTT）** 是一种很强的还原剂，散发一种难闻的气味。可因吸入、咽下或皮肤吸收而危害健康。当使用固体或高浓度储存液时，要戴手套和护目镜，在化学通风橱里进行。

• **溴化乙锭（EB）** 是一种强有力的诱变剂，有毒。有关其特殊的操作和废弃处理程序与当地安全机构联系。避免吸入其粉末。使用含有溴化乙锭的溶液工作时要戴合适的手套。

• **甲酰胺** 是一种致畸物质，其挥发的气体对眼睛、皮肤、黏膜和上呼吸道有刺激

作用。可因吸入、咽下或皮肤吸收而危害健康。当使用高浓度的甲酰胺时，要戴合适的手套和护目镜，在通风橱内进行操作。保持使用液尽可能被遮盖。

• **吲哚乙酸（IAA）** 可因吸入、食入或皮肤吸收而危害健康，并有对眼睛造成严重损伤的危险。使用时，要戴手套和护目镜，在化学通风橱里进行。远离高热、火花和明火。

• **异丙醇** 有刺激性，可因吸入、咽下或皮肤吸收而受害。戴合适的手套和护目镜，切勿吸入其挥发气体。远离高热、火花和明火。

• **β-巯基乙醇（$HOCH_2CH_2SH$）** 如果吸入或通过皮肤吸收可以致命，如果咽下会也是非常有害的。高浓度的巯基乙醇对黏膜、上呼吸道、皮肤和眼睛有非常严重的损伤作用。β-巯基乙醇有非常难闻的味道。要戴合适的手套和护目镜，在化学通风橱内使用。

• **辛醇** 极易燃。要远离高热、火花和明火。辛醇可因吸入、咽下或皮肤吸收而危害健康。使用时要戴合适的手套和护目镜。

• **苯酚** 有很强的毒性和高度腐蚀性，并能引起严重的灼伤。可因吸入、咽下或皮肤吸收而危害健康。使用时要戴合适的手套、防目镜和防护服，在化学通风橱里使用。如果皮肤接触到苯酚，要及时用大量的水冲洗接触部位，并用肥皂和水洗。切记勿用乙醇洗！

• **氢氧化钾（KOH）和甲醇/氢氧化钾（KOH）** 是高毒性化合物，如果误吞食可能是致命的。可因吸入、咽下或皮肤吸收而危害健康。其溶液有腐蚀性，可引起灼伤。操作要非常小心。戴合适的手套和防护眼镜。

• **蛋白酶 K** 是一种刺激物，可因吸入、咽下或皮肤吸收而危害健康。戴合适的手套和护目镜。

• **放射性物质** 当计划涉及使用放射性的实验时，应考虑同位素的物理化学性质（半衰期、放射类型和能量）、放射性物质的化学形态、它的放射浓度（放射性比度）、总量及其化学浓度。只订购和使用所必需的量。当操作放射性物质时，要始终佩戴合适的手套、护目镜，穿实验工作服。无论由工艺装置产生还是由放射性物质源发射的 X 射线和 γ 射线都是很短波长的电磁波，它们可以从放射源各向同性地散发或聚集成束。它们的潜在危险取决于暴露的时间、辐射的强度和使用的波长。要知道合适的防护屏通常是由铅或其他类似的材料制成的。防护屏的厚度依 X 射线和 γ 射线的能量而定。有关合理使用和丢弃放射性物质的更进一步指导准则请与当地安全机构联系。使用放射性同位素后要实行彻底探测检查。常规放射性的一种便利的计算方法可在如下地址中找到：http://graphpad.com/quickcalcs/index.cfm。

• **RNase A** 是一种刺激物，可因吸入、咽下或皮肤吸收而危害健康。戴合适的手套和护目镜，不要吸入其粉末。

• **十二烷基硫酸钠（SDS）** 有很强的毒性、刺激性和对眼睛造成严重损伤的危险，可因吸入、咽下或皮肤吸收而危害健康。戴合适的手套和护目镜。不要吸入其粉末。

• **氢氧化钠（NaOH）和含 NaOH 的溶液** 有很强的毒性和腐蚀性，操作时要格外小心。戴合适的手套和防护面具。所有其他高浓度碱类溶液的操作均应采取类似

方法。

**• 聚乙二醇辛基苯基醚（Triton X-100）** 能引起严重的眼睛刺激和灼伤。可因吸入、咽下或皮肤吸收而受害。戴合适的手套和护目镜。切勿吸入其蒸汽。

**• 二甲苯**　易燃，高浓度的二甲苯可产生麻醉作用。可因吸入、咽下或皮肤吸收而危害健康。戴合适的手套和护目镜，在化学通风橱里使用。远离高热、火花和明火。

# 索　引

［Ｇｅｎｅｒａｌ Ｉｎｆｏｒｍａｔｉｏｎ］
书名＝遗传变异分析实验指南＝ＧＥＮＥＴＩＣ ＶＡＲＩＡＴＩＯＮ Ａ ＬＡＢＯＲＡＴＯＲＹ ＭＡＮＵＡＬ
作者＝（美）Ｍ．Ｐ．韦纳，Ｓ．Ｂ．加布里埃尔，Ｊ．Ｃ．斯蒂芬斯主编
页数＝４４９
ＳＳ号＝１４０７６１１６
ＤＸ号＝
出版日期＝２０１６．０７
出版社＝科学出版社